공학박사학위논문

심미적 시지각에 대한 인지적고찰

Cognitive Analysis of Visual Aesthetic Perception

2015 년 8 월

서울대학교 대학원

협동과정 인지과학 전공

박 태 서

심미적 시지각에 대한 인지적고찰

Cognitive Analysis of Visual Aesthetic Perception

지도교수 장 병 탁

이 논문을 공학박사 학위논문으로 제출함
2015 년 5 월

서울대학교 대학원
협동과정 인지과학 전공
박 태 서

박태서의 박사학위논문을 인준함
2015 년 7 월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

Abstract

Cognitive Analysis of Visual Aesthetic Perception

Tae-Suh Park
Cognitive Science Program
The Graduate School
Seoul National University

Answering for the question “what is beauty?” has emerged as an issue of psychology, neuroscience and computer science during the last decade, after the long history of exploration in the field of philosophy and aesthetics. Especially, in the field of computer science, computational aesthetics pursues implementing an automated aesthetic judgement system based on the low-level features and machine learning techniques, for tangible application such as content recommendation.

In this paper, as an effort of building a computational model of estimating aesthetic value of photos in content recommendation, a hypothesis that surface curvatures of objects and a place in a scene contribute to the estimation is proposed and implemented as a new visual descriptor named Local Slant Cue (LoSC) which represent catching 2.5D information which traditional local descriptors are hard to catch. Experimental results show its comparable performance just with the 30

percent of computational workload of the previous arts. However, comparative study reveals there exist a kind of "glass ceiling" regardless of feature selection, due to a weird attribute of the mediocre samples, which occupy an absolute majority of any given sample group, in machine learning framework.

Observation to the score distributions of the mediocre group leads to the discovery of significantly high variance in consensus level among human raters for the stimuli. For quantitative validation of the observation, skewness-kurtosis map is adopted as a tool of consensus analysis and applied to a massive photo aesthetics dataset consisting of 225,000 samples, followed by the result of showing validated universality of the observation as one of four patterns, which are incompatible with Gaussianity that has been expected so far.

Several computational models of visual aesthetic perception are proposed and tested from the view of how well they explain the observed patterns, finding the comparative advantage of dynamic systems model. As an effort of elaborating the idea of dynamic systems for the aesthetic perception, a new computational model named as DDM4AP (Drift-Diffusion Model for Aesthetic Perception) is proposed regarding visual aesthetic perception as a result of dynamic interaction between "like" factors and "dislike" factors.

While it is concentrating to explain the wide variance in consensus level, the proposed model predicts a significantly longer latency when appreciating photos the mediocre group rather than the good or the bad, regardless of consensus level. Human participant experiments validate the prediction, supporting the model as reflecting important attributes of visual aesthetic perception in human mind.

In conclusion, this study declares computational aesthetics requires new

approaches of machine learning and computer vision considering dynamic interaction between two contrastive factors and selecting training data and features in accordance with such mixed data

Keywords : Cognitive modeling, affect sensing and analysis, computational aesthetics, dynamic systems, consensus analysis

Student Number : 2011-30048

Table of Contents

List of Tables

List of Figures

CHAPTER 1. Introduction

1.1. Background

Studies on visual aesthetics have gained an increasing attention during the last decade in the fields of psychology (Palmer, Schloss, & Sammartino, 2013), neuroscience (Chatterjee, 2011), and computer science (Galanter, 2012; Joshi et al., 2011), concurrently having their own focuses. Considering the interdisciplinary nature of the topic, the previous works of aesthetics studies before the emergence of computational aesthetics are briefly introduced separately for each discipline from philosophy, psychology, and neuroscience as following.

1.1.1. In Philosophy

Traditionally, beauty has been regarded as a topic of the humanities throughout its long history since Greek philosophers. It was not until 1742 that Aesthetics was established as an independent study from philosophy by Baumgarten in Germany where he coined the term from the old Greek word '*Aesthetica'* for differentiating from sense (Shimamura & Palmer, 2012).

For the nature of aesthetic appreciation, various accounts and perspectives from philosophers across eras and countries have led to the five observations as following:

Firstly, aesthetic judgements are not always associated with function or purpose.

For example, people feel beauty when seeing a jewel or galaxy even though it does not offer any practical function for daily life. Plato was the first who regarded beauty as an independent value from goodness while most previous Greek people including Socrates, projecting beauty as something good and appropriate (Griffith & Ferrari, 2000), had not differentiate these values. Kant described it as ”disinterested interest” (Kant, 1952) properly. For explaining such non-functional pleasure from art, Aristotle and Plato proposed a concept of mimesis regarding art as an incomplete imitation of real world or an ideal, while Plato kept negative view on art due to its intrinsic distortion and deterioration (Griffith & Ferrari, 2000) in opposition to Aristotle's view of accepting as a natural form of pleasure (Butcher, 1951).

Secondly, aesthetic judgement is heterogeneous response evoked by various factors. For the diversity of beauty factors, it is evident that people sense beauty everywhere; in nature, people, art, and even mathematical ideas. In the context of the heterogeneity of aesthetic judgement, a Greek teacher of rhetoric Longinus differentiated sublime from beauty in his treatise *On The Sublime* for the first time (Burke, 1796) and motivated Edmund Burke (Burke, 1812) in the 17th century insisting the contrast between the two. Immanuel Kant in the 18th century identified three features – agreeable, good, and beautiful - evoking pleasure and argued aesthetic judgements are based on evaluating the beautiful while regarding sublime as aesthetic

response to nature (Kant, 1952). Furthermore, Freud argued that the "uncanny," the frightening things caused by fear of the unfamiliar, should be considered as a major topic of aesthetics additionally (Freud, Strachey, Cixous, & Dennomé, 1976).

Thirdly, aesthetic judgement is an emotive perception rather than cognition. In the age of Enlightenment, Lord Shaftesbury insisted the idea of "pleasures of an internal sense of beauty" originated from the perceived attribute not requiring any reasoning or evaluation (Gill, 2011). Francis Hutcheson expanded the idea of a reflex sense by categorizing it as the "internal sense" which is different from mere perceptions of sight and hearing which he called as the "external sense" (Hutcheson, 1729).

Fourthly, like color, beauty is not an objective property but a subjective one in beholder's mind. Hume (Hume, 2000), the Enlightenment thinker, pointed out that, like color, beauty is not quality of a subject but a state of mind of its beholder in his hypothesis of Sympathy. Subjectivity is affected by knowledge and experience of beholders, at least in case of art appreciation (Arnheim, 1954).

Lastly, in spite of such subjectivity, there is a common beauty which is agreeable among people because they share similar structure of human mind. Kant asserted the idea of universality that beauty is an innate ideal shared among people (Kant, 1952). Lipps (Lipps, 1935) expanded the account of sympathy (Hume, 2000) to the theory of

empathy asserting there is a common basis among people for feeling beauty (Jahoda, 2005).

1.1.2. In Psychology

Psychologists have transformed aesthetics as a subject of science by applying measurement since the 19th century when Fechner became a founder of experimental aesthetics, by measuring aesthetic judgement quantitatively and insisted that there are several universal factors including golden ratio and contrast for being beautiful to human mind (Fechner, 1876). His account for beauty as pleasure influenced Berlyne (Berlyne, 1971) who insisted that aesthetic pleasure was induced from arousal evoked by novelty, complexity, surprise, ambiguity, heterogeneity, or irregularity, and Arnheim (Arnheim, 1954) who studied the effect of balance, shape, or form to art perception, among others.

One of the recent dominant approaches for explaining beauty in this field is evolutionary aesthetics (Dutton, 2003), a part of evolutionary psychology. Evolutionary aesthetics assumes that people experience beauty when perceiving something that helped survival of our ancestors. From the view aesthetic perspective, evolutionary psychologists have three notions:

"First, beauty serves as a proxy for health and vigor in mate selection. Second, beautiful objects are those that are complex and yet are processed efficiently. And third,

art making and appreciation serves an important ritualistic function that enhances social cohesion." (Chatterjee, 2011)

For example, several researchers explained facial beauty as the result of evolutionary adaption for finding a healthier mate based on certain physical features of faces and bodies such as symmetry, bright skin color, and hairs. Natural selection for survival and sexual selection for reproduction are the two underpinnings; Miller (G. F. Miller, 1999) insists sexual selection more than natural selection explains many human indulgence such as arts and culture.

Facial attractiveness is the most representative case of showing the consistency of its ratings across ethnicities and cultures (Cunningham, Roberts, Barbee, Druen, & Wu, 1995; Langlois et al., 2000). For example, larger eyes and delicate jaws are common factors of attractiveness regardless of culture or ethnicity. Other factors including Averageness (Rubenstein, Kalakanis, & Langlois, 1999; Thornhill & Gangestad, 1999), symmetry (Grammer & Thornhill, 1994; Mealey, Bridgstock, & Townsend, 1999), sexual dimorphism (Buss, 1989; Grammer, Fink, Møller, & Thornhill, 2003; Perrett, May, & Yoshikawa, 1994) have been proposed as the beauty factor of faces. Following interpretation of evolutionary psychologists, the preference to an average face reflects the innate bias to majority or prototype (Mervis & Rosch, 1981), and symmetry is utilized as a sign of healthy nervous system and thereby choosing a better

mate. Such a point of view of regarding facial beauty as the sign of the fittest can be supported by the research showing that emphasis on physical attractiveness is highly correlated with infestations of malaria and parasites across human societies (Gangestad & Buss, 1993).

Also, many experimental results support a hypothesis that preference to facial beauty is established in very early stage of development: infants play with a more attractive doll longer than the less one (Langlois, Roggman, & Rieser-Danner, 1990), and stare at the beautiful faces longer while the beautiful faces are selected from the view of adults(Langlois, Ritter, Roggman, & Vaughn, 1991; Slater et al., 1998). Interestingly, the facial attractiveness can be moderated by locational context, at least when female raters evaluate male subjects. Specifically, preference to male beauty is weaken when the photo is provided with negative description about the place as dangerous and full of criminals(Leder, Tinio, Fuchs, & Bohrn, 2010).

Although it is hardly exposed, bodily attractiveness is also affected by averageness symmetry, and sexual dimorphism: symmetry (Grammer & Thornhill, 1994; Møller, 1992; Møller & Swaddle, 1997; Thornhill & Gangestad, 1994), height (Jackson & Ervin, 1992), broad shoulders of male mates (Horvath, 1979), among others. Even though it is affected by cultural context named as "*environmental security hypothesis*" (Pettijohn Ii & Tesser, 1999; Pettijohn & Jungeberg, 2004), bodily

attractiveness of female mates is dominated by a universal feature of an hour-glass shape in 0.7 of a waist-to-hip ratio (D. Singh, 1993). Contrary to the case of a face, bodily attractiveness can be enhanced by movement like dancing (Singer et al., 2000) or dimorphic walking patterns (M. P. Provost, Quinsey, & Troje, 2008).

Scene beauty was the main topic discussed among aesthetic theorists in the 18th century: preferring natural scenes to artificial scenes is the representative example (S. Kaplan, Kaplan, & Wendt, 1972). In case of scenes, from the view of hunter-gather ancestors, safety and rich resources should have been good criteria and then hardwired to human brains as a beauty instinct. as place improves the ancestor's chances of survival in the Pleistocene era, from 1.8 million to about 10,000 years ago: "the savanna hypothesis"(Balling & Falk, 1982). Water, large trees, a focal point, and open spaces have been proposed as the factors for judging scene aesthetics (Han, 2010; R. Kaplan, Kaplan, & Brown, 1989; Orians & Heerwagen, 1992).

However, not all aesthetic experiences are tied to something useful. Especially in art, "disinterested interest" (Kant, 1952) is self-contained and not tied to something useful while it can be interpreted as a "spandrel"(Gould & Lewontin, 1979), an unintended by-product of evolution, although a few researchers with Darwinian perspective suggested the benefit of recognizing several aesthetically pleasing and displeasing factors in preying or mating (Dutton, 2003). Scherer (Scherer, 2005)

pointed out, without aesthetic emotions, utilitarian emotions are not enough to explain affective response of human being. Another challenges concerning appreciation of art are subjectivity and inconsistency. The link between specific percepts and the reward system changes based on knowledge, experience, and the health status (Chatterjee, 2011). A vivid divide in the response to contemporary arts between trained critics and laymen might be the consequence of its property of being less-dependent to utility-based preference engine in a human brain and effect of experience and knowledge.

Lastly, psychologists have repeatedly proven the importance of understanding beauty in commercial applications: e.g., facial beauty seems affecting teachers evaluating students (Kenealy, Frude, & Shaw, 1988), pedestrians trying to return a wallet they picked on the street (Sroufe, Chaikin, Cook, & Freeman, 1976), people receiving a misdelivered post (Benson, Karabenick, & Lerner, 1976).

1.1.3. In Neuroscience

Owing to the development of in-vivo brain imaging technologies including fMRI, neuroscientists started looking into the human brain when appreciating art, notifying the coming of neuroaesthetics (Chatterjee, 2011). Neuroaesthetics have three notions (Chatterjee, 2004) as following: firstly, visual aesthetics have multiple components like other vision functions; secondly, aesthetic experience emerges from the combinations

of the responses from the components; thirdly, visual aesthetics should reflect the hierarchical sequence of visual processing along early, intermediate, and late vision as Marr defined (Marr, 1982). Some neuroscientists count context and cultural factors additionally (Jacobsen, 2006; Leder, Belke, Oeberst, & Augustin, 2004).

Because they share the concept of aesthetic judgement from brain activity and of brain as the product of evolution as the basic assumption, neuroaesthetics and evolutionary psychology are regarded as explaining how and why of aesthetics. Especially, facial beauty appreciation has been popular in both domains owing to its universal and innate nature (Jones & Hill, 1993; Langlois et al., 2000; Langlois et al., 1991; Perrett et al., 1994; Slater et al., 1998) although there is a cultural variation (Cunningham, Barbee, & Philhower, 2002) to some extent.

Zeki (Zeki, 1999), the pioneer of neuroaesthetics, proposed parallelism between art and neuroscience as the neural networks extract attributes like color, brightness, and motion from visual stimuli while artists transform the attributes further (Ungerleider, 1982) (Conway & Livingstone, 2007) and pursue perceived distortion rather than physical accuracy in painting (Cavanagh, 2005). In the same vein, Ramachandran & Hirstein insisted that artists empirically find out visual primitive which evoke peak response from a set of perceptual principles (Ramachandran & Hirstein, 1999).

Early studies in this field concentrated on the case of artists who damaged in their

brains (Chatterjee, 2006; Chatterjee, Hamilton, & Amorapanth, 2006; Zaidel, 2005). For example, some patients with fronto-temporal dementias (FTD), a kind of obsessive-compulsive disorders, show significant enhancement of artistic performance while they suffer from disorganization and deficiency of attention, linguistic ability, and decision making (B. L. Miller & Hou, 2004), let alone the case of autism (Sacks, 1995). The patients usually have damages in orbito-frontal and medial-temporal cortices and fronto-striatal circuits while posterior occipito-temporal cortices, the core part for object and location recognition, are intact (Kwon et al., 2003; Ursu, Stenger, Shear, Jones, & Carter, 2003). Even a framework was proposed for measuring the components of art and thereby finding the correlation with the brain damage of the artists (Chatterjee, Widick, Sternschein, Smith, & Bromberger, 2010). A more natural setting of observing brain activities during aesthetic judgement have followed and compensated the above researches.

Experimental results from the in-vivo brain imaging research on art appreciation imply that decision making and emotional reward involve aesthetic judgement. In their pioneering work, Kawabata and Zeki (Kawabata & Zeki, 2004) showed that the orbitofrontal cortex is activated during aesthetic judgement which implies it is similar with decision making rather than sensing or mere perception. The reward system also involves in the process; for example, beautiful faces activates the reward system

including the ventral striatum, the orbito-frontal cortex, and the nucleus accumbens, (Aharon et al., 2001; Ishai, Fairhall, & Pepperell, 2007; Kampe, Frith, Dolan, & Frith, 2001; Kranz & Ishai, 2006; O'Doherty et al., 2003), the amygdala(Winston, O'Doherty, Kilner, Perrett, & Dolan, 2007), related with emotional valences (Senior, 2003). Furthermore, some researchers (Chatterjee, Thomas, Smith, & Aguirre, 2009) believe that ventral occipital region is responsible for automatic judgement of aesthetics and perceived beauty-virtue proximity that attractiveness of a person evoke biases in the estimation of the person's intelligence, honesty, leadership (Kenealy et al., 1988; Lerner et al., 1991; Ritts, Patterson, & Tubbs, 1992), and strength (Dion, Berscheid, & Walster, 1972). In case of landscape appreciation, a beautiful place reportedly evoked more activation in the right side than the left side of parahippocampus (Yue, Vessel, & Biederman, 2007).

However, a careful interpretation of brain imaging result in neuroaesthetics is required because it is hard to differentiate the activation of a specific brain region by cognitive process from that by aesthetic judgement because the two processes run together automatically regardless of the task given to human participants (Chatterjee et al., 2009). For an instance, seeing portrait, still life, landscape accompany relatively high activation in the lateral occipial cortex (LOC), the fusiform gyrus (FFA), and the parahippocampus (PPA), respectively because these regions are responsible for

information processing of respective subjects, not for aesthetic judgement, although there is a report that a more beautiful face evokes relatively higher level of activation in the face area (FFA) and the object area (LOC)(Chatterjee et al., 2009). Brain imaging researches for finding counterparts of appreciating beauty have produced inconsistent and various reports (Chatterjee, 2011). First of all, as aesthetic judgement accompanies emotional response, it is natural for the reward system such as the medial and orbito-frontal cortices, the anterior medial temporal lobe, and the subcortical structures involve the process (Berridge & Kringelbach, 2008; O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001). The first fMRI imaging experiments for subjects judging aesthetic classes (Kawabata & Zeki, 2004) reported activation of orbito-frontal cortex (BA 11) during the "good vs bad" task, and that of left parietal cortex (BA 39) and anterior cingulate (BA 32) during the "good vs neutral" task. They also reported that, across four types of stimuli consisting of portrait, landscape, still image, and abstract art, only the orbito-frontal cortex showed consistent activation for "good" stimuli, insisting that the region is the neural correlate of feeling "goodness." Another 2-class study using magnetoencephalography pointed out the left dorsolateral prefrontal cortex as the region of aesthetic judgement, based on the observation of strong activation around 400-1000msec for the task of "good vs not good" (Camilo J. Cela-Conde et al., 2004). Studies using Likert scale regression rather than the above

classification, the left anterior cingulate and occipital gyri and were reportedly activated (Vartanian & Goel, 2004); the right caudate activated proportionally to the score. The fMRI study of appreciating geometric diagrams showed preference to symmetry and activation in the precuneus, medial frontal cortex, and ventral prefrontal cortex around 360-1225msec (Jacobsen, 2006).

In conclusion, providing a comprehensive model of explaining such various results is required (Biederman & Vessel, 2006), including a general model combining the reward systems and the decision making systems (Chatterjee, 2004; Nadal, Munar, Capó, Rossello, & Cela-Conde, 2008) which has yet to come.

1.2. Related Works: Computational Aesthetics

Computer scientists are the latest group of joining the study of beauty following philosophers, psychologists and neuroscientists as discussed above briefly. The research named "computational aesthetics" (Joshi et al., 2011)is motivated by the need for computation model which is able to estimate digital contents such as photos, videos, or songs as a part of content-based recommendation system (CBRS)(Ricci, Rokach, & Shapira, 2011). Specifically for photo recommendation, this feature plays an important role of selecting the top 5 (or 10) items from millions of candidates to be provided in the first page of search result, the very place where users evaluate quality of

recommendation. The primitive selection scheme based on image quality is becoming insufficient to select the top items as quality of pictures has standardized upward according to the development of digital photography technologies, calling for a more elaborated estimation engine based on aesthetic quality (Joshi et al., 2011) additionally. For the purpose, data-driven analysis of image aesthetics has been proven as useful for elaborating the search result from massive photo collection (Obrador, Schmidt-Hackenberg, & Oliver, 2010) or even suggesting the best scenic angle for amateur photographers (Su, Chen, Kao, Hsu, & Chien, 2012), for example.

This research on data-driven computation aesthetics has also been conducted as a part of content-based recommendation system named "CogTV" at Seoul National University where multimodal cues are used to construct a computational model of being trained by a TV viewer's non-verbal physiological feedback, the signal of emotional response related with satisfaction, in real-time (T.-S. Park, Kim, & Zhang, 2014), according to the tradition of affective computing (Hanjalic & Xu, 2001; Picard, 1997). Its early result of underlining the importance of visual stimuli compared with the other modalities like audio and text resulted in the further research of seeking beauty factor beyond the mere quality of images.

Feature selection is one of the most important parts for making computational models of visual aesthetic perception. For the issue, previous researches in the field

of psychology and neuroscience suggest various "aesthetically pleasing" or "preferred" factors in images including color (Guilford & Smith, 1959; McManusU, Jones, & Cottrell, 1981; Ou, Luo, Woodcock, & Wright, 2004; Palmer & Schloss, 2010), curved object(Bar & Neta, 2006; Leder, Tinio, & Bar, 2011), contour (Vartanian et al., 2013), canonical size (Konkle & Oliva, 2011; Linsen, Leyssen, Gardner, & Palmer, 2010), and spatial composition (Leyssen, Linsen, Sammartino, & Palmer, 2012; Sammartino & Palmer, 2012) (see (Peters, 2007) for categorized summary). In computer science, Savakis and his colleagues took a pioneering step by performing a large scale study of possible features which affect aesthetic rating, reporting several semantic and style-related features (e.g., composition, baby, colorfulness) as significant factors (Savakis, Etz, & Loui, 2000).

In the same vein, since Datta et al.(Datta, Joshi, Li, & Wang, 2006), most researchers in early computational aesthetics have focused on implementing the empirical heuristics of painters and professional photographers such as composition (Obrador et al., 2010), people and vanishing points (Cerosaletti & Loui, 2009), or rule of thirds (Mai, Le, Niu, & Liu, 2011), in addition to visual weight balance, high dynamic range, and contrast color harmony; for more information, see the reviews(Galanter, 2012; Joshi et al., 2011).

In addition to the heuristic approaches, several researchers have shown the

possibility that generic low-level features such as color histogram or GIST (Oliva & Torralba, 2006) can be effective for estimating the visual aesthetic value of a photo directly in the framework of machine learning (Datta et al., 2006; Marchesotti, Perronnin, Larlus, & Csurka, 2011). Such usage of low-level features is justified by several neuroscientific discoveries (Ishai et al., 2007; Russell & George, 1990; Woods, 1991), showing that aesthetic perception treats form and content separately; as the early and intermediate vision process form while the late vision is for content, using the low-level features might explain the effect of form, at least. Recently, Su et al.(Su et al., 2012) reported that classifier construction using bottom-up-induced features reproduces the major two heuristics, visual weight balance and a rule of thirds, in its characteristics. Considering the practical limitation that not all high-level features are learnable in machine learning scheme and the unresolved fundamental issue of whether or not any high-level feature is deterministic to evaluate perceived beauty, using low-level features for learning aesthetic values of photos has merit of reflecting not only explicit beauty factors (e.g., a rule of thirds) but also implicit and unrecognized factors owing to its explorative search power, as far as the machine learning result converges to a global optimum sufficiently.

Combined with appropriate features, a reliable and representative dataset is essential for training an aesthetic value estimator in machine learning approach. Owing

to the Internet and digital photography, in the field of affective computing, crowdsourcing has become a popular method of gathering massive data of self-reported scores about visual stimuli via interactive environment such as Mechanical Turk (e.g., see (Soleymani & Larson, 2010)) or photography-dedicated web sites like http://www.dpchallenge.com or http://www.photo.net. Based on the two web sites, several datasets have been proposed as a reference including Photo.net (Datta et al., 2006), CUHK dataset (Ke, Tang, & Jing, 2006), and AVA dataset (Murray, Marchesotti, & Perronnin, 2012). The latest one, AVA (an acronym for Aesthetic Visual Analysis), consists of 255,800 photos with 10-point-scale aesthetic values annotated from hundreds to thousands users of DPChallenge.net. For 29,451 photos, it provides at least one semantic or style tag selected from 65 categories. Owing to its large scale and well-documented properties, the dataset is gaining popularity among computational aesthetics researchers.

Compared with the numerous factor analyses, relatively a fewer conceptual models (Leder et al., 2004; Reber, Schwarz, & Winkielman, 2004) have been proposed to explain the effect of such factors to aesthetic perception.

In conclusion, the current computational aesthetics have not fully leveraged aesthetics studies accumulated in science and humanities, although it partially depends on a few heuristics derived from the studies. Specifically, there are two approaches of

comprising the unused lessons from the studies: spatial composition should be captured by all means in feature space; and, multi-factors invoking emotional rewards should be a part of the process of aesthetic appreciation. In the following chapters, Chapter II tackles the first approach by designing a feature for capturing spatial layout efficiently, while the second approach is implemented in Chapter IV as a fundamental assumption of a new model to be proposed.

CHAPTER 2. Finding Features

2.1. Background

In the current machine learning approaches with low-level features for computational aesthetics, it is hard to see dominance of a specific feature or descriptor for the purpose. The representative features designed for object recognition or scene classification including SIFT (Lowe, 2004), GIST (Oliva & Torralba, 2006), and HoG (Dalal & Triggs, 2005) have been used widely in this field. In usual implementations, SIFT (Lowe, 2004) or HoG (Dalal & Triggs, 2005) have been used with the bag of visual words as the combination provided competitive performances with a relatively short feature vector of 128 dimension. However, as Wu and Rehg (J. Wu & Rehg, 2011) pointed out, both SIFT (Lowe, 2004) and bag of word lack spatial structure and generalizability, while HoG (Dalal & Triggs, 2005) relatively preserve them just in edge space. GIST(Oliva & Torralba, 2006), on the other hand, is designed to represent spatial layout and is known to be good at recognizing natural scenes while it seems insufficient to capture properties of objects or indoor scenes (J. Wu & Rehg, 2011). Considering such characteristics, the combined feature set has been proposed: e.g., SentiBank (Borth, Chen, Ji, & Chang) consists of RGB histogram, GIST, LBP, and Bag of Words.

The importance of spatial information for aesthetic judgment has been supported

by various researches (Cerosaletti & Loui, 2009; Ciesielski, Barile, & Trist, 2013; Obrador et al., 2010; Savakis et al., 2000). Aligned with the report from Savakis et al.(Savakis et al., 2000), Obrador et al. pointed composition as a key factor of scene beauty, consisting of simplicity, balance, and geometry, and achieved 66.5% accuracy in discriminating 538 best-worst photos by capturing simplicity with number of colors, proportion of big segments, and out-focus while checking rule of thirds and golden ratio(Obrador et al., 2010).

Using spatial information for aesthetic judgement has merit in minimizing the effect of subjectivity by concentrating on the early stage of visual aesthetic appreciation rather than the late one. Spatial layout has been known one of the earliest percepts (Oliva & Schyns, 1997) and therefore expected to be appreciated early, meaning that it is relatively less affected by late semantic cues or memories, the very factors contributing to the subjectivity issue in aesthetic appreciation (Arnheim, 1954; Chatterjee, 2011). Therefore, using visual stimuli consisting of space-centric subjects (e.g., landscape or cityscape) is believed to be effective for concentrating on early appreciation of aesthetic value and thereby control subjectivity to some degree. This assumption is also supported by evolutionary psychologists reporting more consistency among raters in natural scenes rather than abstract or artificial subjects (Dutton, 2003).

Another research reports presence of people, perspective cues , and size of the

main subject are major beauty factors with three-way interaction among them(Cerosaletti & Loui, 2009). They found several interesting patterns in their extensive study including: presence of people in a photo is the strongest factor; perspective is important only if the main subject is small; facial expression matters if exist; and horizontal cue is far more preferred to upward view, among others(Cerosaletti & Loui, 2009).

In addition, it is unique to computational aesthetics that quality features have also been used broadly such as hue, dark channel prior (He, Sun, & Tang, 2011), saturation(Ciesielski et al., 2013) and sharpness. Cerosaletti and Loui (Cerosaletti & Loui, 2009) also validated the correlation between poor photo quality and low aesthetic perception in their experiment: see (Bhattacharya et al., 2013; Bhattacharya, Sukthankar, & Shah, 2011; Ke et al., 2006) for example.

Because visual aesthetic perception is specific to neither objects (e.g., portrait and still life) nor scene (e.g., landscape), a feature or descriptor for computational aesthetics should represent both categories well, as complying with the conditions for designing good visual descriptors (Mikolajczyk & Schmid, 2005). Several psychological studies also support the idea of contribution of spatial information to aesthetic evaluation including canonical size (Konkle & Oliva, 2011; Linsen et al., 2010) and spatial composition (Leyssen et al., 2012; Obrador et al., 2010; Sammartino & Palmer,

2012). For spatial layout of regions, although the visual word model usually ignores it, several works leverage it as a key element: Spatial pyramid matching (SPM) (Lazebnik, Schmid, & Ponce, 2006) and CENTRIST (J. Wu & Rehg, 2011), among others.

CENTRIST (J. Wu & Rehg, 2011) is regarded as one of the current state of the art for scene description considering its nonparametric nature, cross-class performance, and designed merit for implementing fast feature extractors. It utilizes histogram of Census Transform (Zabih & Woodfill, 1994), CT, and thereby inherits the merits of CT as a robust descriptor to illumination change. CENTRIST is also successful by avoiding common misuse of CT values as integers which is comparable in amount by converting to a histogram that its number of bins is smaller than 256; considering their nature, CT or LBP (Ojala, Pietikainen, & Maenpaa, 2002) values should be treated as a texture pattern descriptor that all bits are independent to each other and therefore treated as a kind of texture descriptor, as shown in several articles. One bit difference between two CT values does not guarantee proportional difference to the single bit change between them, which is natural in numeric relation.

Another missing property so far that a good feature or descriptor for computational aesthetics should capture is curvature. Since Burke (Burke, 1812) mentioned smooth curvature as one of the beauty factors, smoothness in curved

object(Bar & Neta, 2006; Leder et al., 2011) or contour (Vartanian et al., 2013) have been reported as a significant factor for being evaluated as beautiful. Datta (Datta et al., 2006) proposed a feature for measuring shape convexity but didn't show significant enhancement.

2.2. Local Slant Cue (LoSC)

The lessons about the ideal features for capturing aesthetic factors leads to an idea that a good aesthetic descriptor for objects and scenes should represent the three dimensional spatial characteristics of surroundings, which has been essential for evaluating the feasibility of ego-motion in any given direction since the birth of the first animal with eyes, from the information of surface distribution rather than edge distribution, because edges often have different meanings such as a border and texture on the flat surface.

Although the idea of emphasizing on surface as the spatial cue has a long history (S. Singh, Gupta, & Efros, 2012) including Marr (Marr, 1982) and Gibson (Gibson, 1950), its implemented descriptors (e.g., gradient image) have not gained popularity as much as SIFT (Lowe, 2004) or HoG (Dalal & Triggs, 2005) due to vulnerability to illumination change or textured surface. To overcome the weakness of the previous approaches while describing more than two dimensional world under any environment,

I propose a novel binary descriptor, local slant cue (LoSC), derived from the philosophy of seeing space as a collection of local tiny surfaces which contribute to spatial perception only in collective manner.

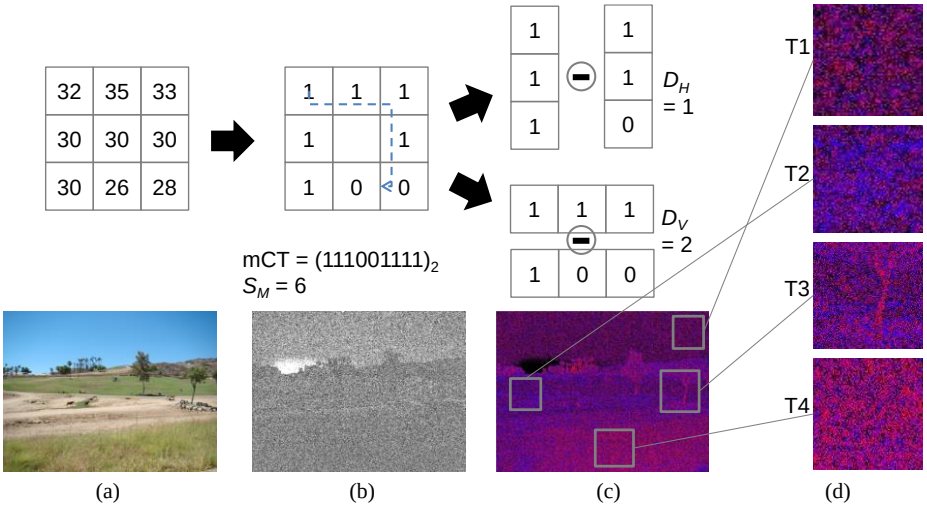


Figure 1. Representation of Local Slant Cue (LoSC) in pixel level

2.2.1. Representation

Inspired by Census Transform (CT) (Zabih & Woodfill, 1994), LoSC converts luminance values of an image into relative order between center and its surround as binary patterns with clockwise order, as depicted in Figure 1, while CT assumes the order of bits in symmetric manner across horizontal and vertical axes.

As the first step, LoSC extracts a modified CT values for all 3x3 pixel regions. The original CT value is calculated as following:

$$CT = \sum_{p=0}^{7} s(I_c - I_p)\, 2^p,\ s(t) = \begin{cases} 1\ if\ t \geq 0 \\ 0\ if\ t < 0 \end{cases} \quad (1)$$

For the technical reason, LoSC changed the sequence of *Ip* in Equation 1: the

upper left is set to the origin (p = 0) and the *p* increases in clockwise manner which is depicted as the dashed arrow in Figure 1 and Equation 2 below. It also extends to 9-bit 1-D array by filling the most significant bit (MSB) with the least significant bit (LSB). Such modification helps keeping the numeric relation between two similar CT values from the view of not only the amount of set bits but their proximity to each other. Even though the original CT helps preserving the correlation between two adjacent 3x3 regions, LoSC regards preserving the invariance to a small change in a tiny region is more important even though it costs the CT's original merit.

$$mCT_i = \begin{cases} CT_i \; if \; i < 8 \\ CT_0 \; if \; i = 8 \end{cases} \tag{2}$$

The two components for each pixel in LoSC representation, S_M and S_D, are calculated by bit operations as following:

$$S_M(x,y) = popcnt\big(mCT(x,y)\big) \tag{3}$$

$$S_D(x,y) = \{D_H(x,y), D_V(x,y)\}$$

where

$$D_H = \left| \sum_{i=2}^{4} mCT_i(x,y) - \sum_{i=6}^{8} mCT_i(x,y) \right| \tag{4}$$

$$D_V = \left| \sum_{i=0}^{2} mCT_i(x,y) - \sum_{i=4}^{6} mCT_i(x,y) \right|.$$

The directional components consisting of the angular image S_D can be visualized

intuitively by rendering D_H and D_V as red and blue respectively. A vivid bluish or reddish region implies there are enough spatial cues in the patch; contrarily, a relatively dark region implies it lacks spatial layout information and therefore it is likely to be a part of void (T1 in Figure 1) or dark regions. If a region looks vivid without single color dominance between blue and red, it means that the region is likely to contain non-directional textures. Also, none of LoSC representation is proportional to illumination or exposure nor dependent to chromatic information, even though there is no technical reason preventing from using it to chromatic channels.

As explained in the first part of this chapter, more general information on spatial layout seem at the slant of local surfaces, while CENTRIST (J. Wu & Rehg, 2011) regard edges as also having information about spatial layout.

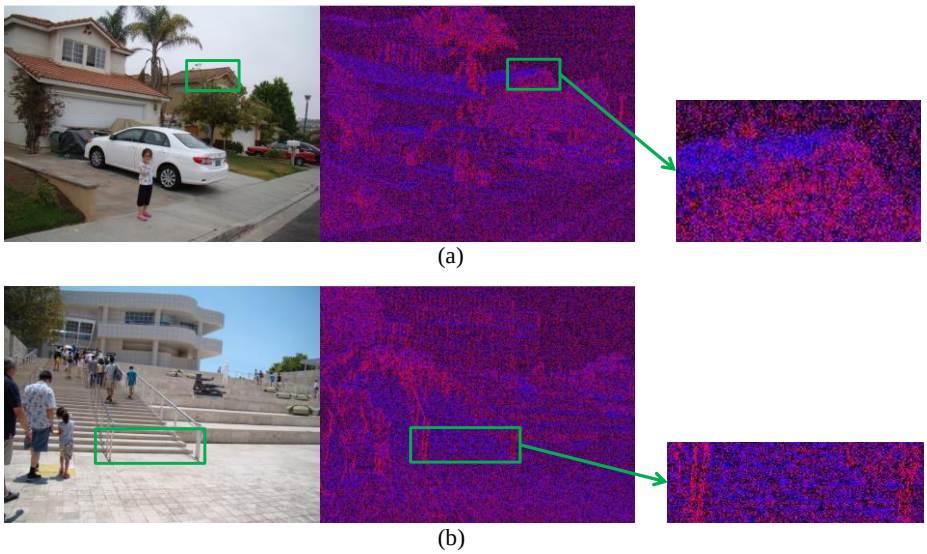


Figure 2. Example of LoSC representation in urban layout

2.2.2. Region Description

Instead of using the popular spatial pyramid matching (SPM) (Lazebnik et al., 2006) scheme or spatial PACT (J. Wu & Rehg, 2011), a simple grid scheme with different dimensions (1 by 1, 2 by 2, 1 by 3, 3 by 1, 3 by 3, and 4 by 4) is adopted to reflect spatial openness; for example, a spatial layout of typical open space consisting of sky, subjects, and ground is expected to be better captured by the "3 by 1" region template. Owing to the naive layout and independence between adjacent 3x3 blocks, contrary to CT and CENTRIST, LoSC is robust to rotation except upside down as it usually affects one of the fundamental assumption of sky-ground relation.

For each region of interest, the LoSC descriptor represent it as a pair of R_M and R_D, which are used directly as feature elements for machine learning.

$$R_M = \frac{1}{n}\sum_{i=0}^{n-1} S_M(i) \tag{5}$$

$$R_D = C_R\left(\frac{D_H - D_V}{2(D_H + D_V + \varepsilon)}\right) + 0.5 \tag{6}$$

where the confidence coefficient C_R is calculated as following:

$$C_R = \frac{1}{n}\left(\sum_R f(D_H) + \sum_R f(D_H)\right) \tag{7}$$

$$f(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

Theoretically, an arbitrary region of interest with any shape can be converted into two scalar values of R_M and R_D. This scheme even comprises both a relatively flat surface with 3D rotation (Figure 2a) and a complicated spatial structure such as stairs (Figure 2b).

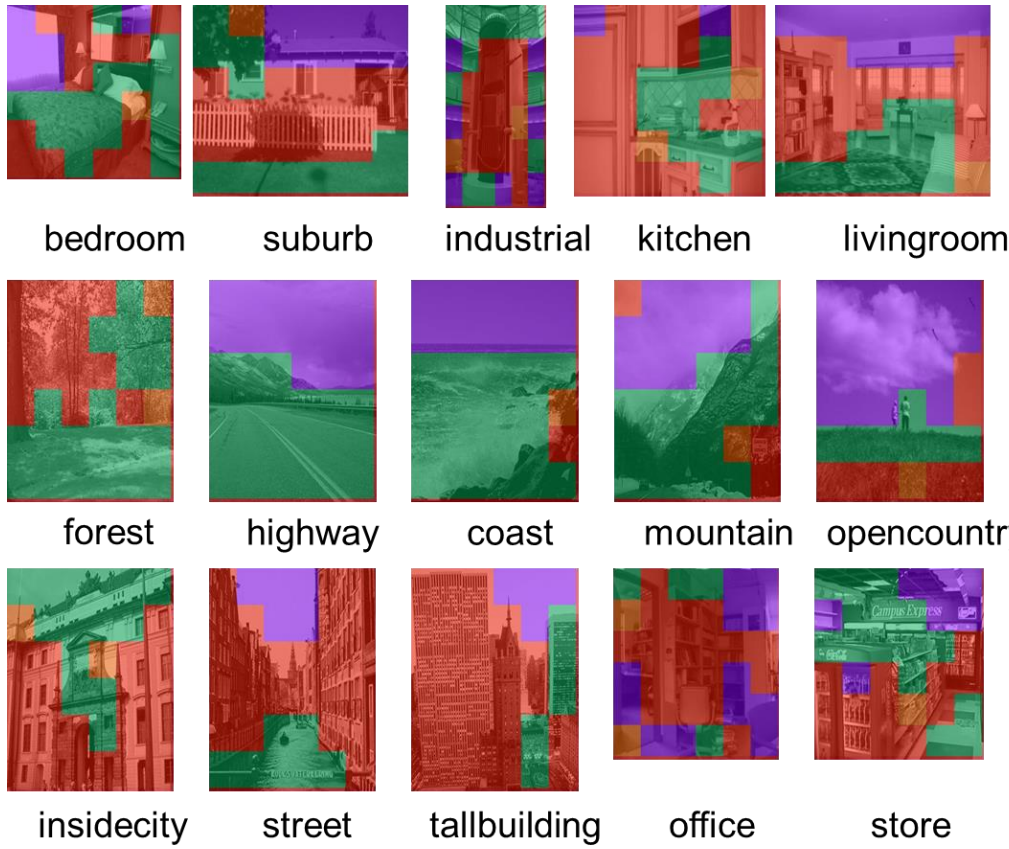


Figure 3. Examples of three geometric class representation based on LoSC. Sky is rendered as violet, wall as red, ground as green.

Figure 3 shows the case where LoSC is used for classifying three geometric classes (sky, wall, ground) just by combining with a set of conditional decisions and thereby revealing the potential merit of LoSC as a slant descriptor.

Although a simple grid map seems sufficing the intended tasks, it deserves to consider combining LoSC with the emerging concept of mid-level discriminant patch by unsupervised manner (Juneja, Vedaldi, Jawahar, & Zisserman, 2013; Xiao, Hays,

Ehinger, Oliva, & Torralba, 2010).

Lastly, workload analysis reported that average time of generating LoSC descriptors for an image of VGA resolution (640x480) was measured as 168 millisecond at Windows 7 PC running on Intel i7 3770 3.40GHz processor with 8GB Ram, which is approximately a third of workload for extracting common combination of SIFT and GIST.

2.3. Experiments

For aesthetics categorization, the latest AVA dataset (Murray et al. (Murray, Marchesotti, & Perronnin, 2012) was used for acquiring photos and their tags. Since an aesthetic score was rated in 10-point scale for each photo, following the guideline of the dataset, the scores were classified as bad, mediocre, and good by calculating mean and standard deviation of mean scores for each tag group and by using (mean-1σ) and (mean+1σ) as two thresholds between the three classes.

A feature vector of LoSC consists of R_M and R_D as a real-number array with 136 dimensions. Another major feature, Hue and Tone (HT) signature (Kobayashi, 1981), was also adopted as it has been used as an emotion research in affective computing (Solli & Lenz, 2010). In addition to LoSC and HT color, six quality features (Weber contrast, sharpness, three color dynamic ranges, and number of pixels) and three

saliency features were added, resulting in up to 1291 dimensions.

Training is performed using WEKA (Hall et al., 2009) with random forests (Breiman, 2001), an SVM-variant (SMO) Platt 1999 (Platt, 1999) with RBF kernel, and AdaBoost with J48 decision tree. All parameters were optimized via grid search. In case of severely imbalanced samples, SMOTE (Chawla et al., 2002) and spread subsampling was applied as preprocessing.

Table 1 shows the comparative performance result from a pilot test for choosing appropriate classification methods for the AVA dataset. The numeric scores in the table do not matter because the test used the training set for qualitative analysis of feature space rather than finding optimal configuration. The low performance of Naïve Bayes or its AdaBoost version implies that the feature space is complicated and highly correlated among features, confirming Random Forest, SMO, and kNN as the major methods for classification hereafter.

Table 2 shows the performance results of optimized good-versus-bad classifiers for five representative photo groups (two for scene-oriented, the other three for object-oriented) with four different classifiers. In case of using HT colors and LoSC as features and SMO (with an RBF kernel) as a classifier, the pictorial aesthetic class estimator achieved 72.36% for landscape, 65.97% for cityscape, 59.73% for portrait, 65.80% for animal, and 70.18% for floral in accuracy with the best case. Compared

with SMO, Random Forest with 160 trees showed relatively stable performance across subjects, including a better performance of 62.60% for portrait. In general, the performance of the estimator was comparable to the reference performance of 65~67% given from AVA dataset authors (Murray, Marchesotti, Perronnin, & Meylan, 2012) achieved by the combination of SIFT (Lowe, 2004), color, and Gaussian Mixture Model.

However, when it came to three-class problem including the mediocre group, the classifiers suffered from the serious imbalance and innate properties as shown in Table 3. In the table, all cases were tested using raw – therefore imbalanced – test data while adopting different training strategies: the "raw" case used another imbalanced training data; the "adjusted weight" case controlled the learning rate according to the proportion of each group in size during training; and, the "rebalanced" group used a rebalanced training data by subsampling. Because accuracy used in Table 1 and Table 2 were inappropriate to describe the imbalance problem properly, mAP (mean Average Precision) was used instead in Table 3. For all cases, inclusion of the mediocre group significantly distorted the training of both classifiers: (a) "RF200" represents Random Forest with 200 trees while (b) "SMO_C5G0.01" stands for SMO with 5.0 of C (the margin parameter) and 0.01 of the gamma. Compared with the two-class case, net performance plummeted once including the mediocre group owing to its intrinsic

dominance in proportion of data. It is interesting that enhancement by rebalancing was limited, implying that it is more than the matter of proportion.

Table 1. Comparative two-class (good, bad) result from a pilot test for choosing appropriate classifiers in AVA dataset: training set is used for testing.

Method	Option	Accuracy	ROC	Precision	Recall
NaiveBayes	N/A	64.66	0.71	0.7	0.51
SMO	-C 5.0 -G 1.0	96.77	0.97	0.98	0.96
kNN	k=31	89.34	0.89	0.98	0.8
J48	-C 0.01	83.56	0.84	0.85	0.82
RandomForest	-I 30	92.53	0.98	0.93	0.91
AdaBoost	NaiveBayes	70.1	0.76	0.76	0.59

Table 2. Binary classification performance (accuracy in percent) for five photo groups: 3-fold cross validated. HT Color and LoSC are used.

	Landscape	Floral	Cityscape	Animal	Portrait
NaiveBayes	61.80	61.37	61.39	56.77	55.81
kNN	67.20	63.92	57.50	58.01	54.20
Random Forest	68.67	66.59	64.31	66.82	62.60
SMO w/ RBF	72.36	70.18	65.97	65.80	59.73

Table 3. Three-class classification performance for three training strategies

Training Set	Precision-Good	Precision-Mediocre	Precision-Bad	mAP
Raw	0.00	0.65	0.50	0.38
Adjusted weight	0.53	0.65	0.25	0.48
Rebalanced	0.27	0.70	0.26	0.41

(a) RF200

Training Set	Precision-Good	Precision-Mediocre	Precision-Bad	mAP
Raw	1.00	0.65	0.00	0.55
Adjusted weight	0.30	0.67	0.29	0.42
Rebalanced	0.25	0.69	0.25	0.40

(b) SMO_C5G0.01

2.4. Discussion

The proposed new low-level visual descriptor, Local Slant Cue (LoSC), is dedicated to capture slants regardless of surface texture, based on the concept of representing a spatial layout of scene as regionally distributed local slants. The new feature inherits all benefits, including the robustness to illumination change like CENTRIST (J. Wu & Rehg, 2011) or Local Binary Pattern (Ojala et al., 2002) while focusing on surface information rather than shape information via edges.

However, the degree of enhancement is not so significant according to the choice of descriptors, limited to less than 70 percent in accuracy for 2-class categorization and worse for the 3-class case in AVA dataset. The toughest issue from the view of feasibility in practical application is in the mediocre group: the accuracy of the mediocre group was limited to 33 percent in its best setting with balanced dataset. Considering the tendency that the mediocre group occupies majority in a real application (in other words, evidently good or bad samples are rare among the randomly chosen samples), it is evident that the weakness is responsible for deterioration of net performance when applied to real-world application.

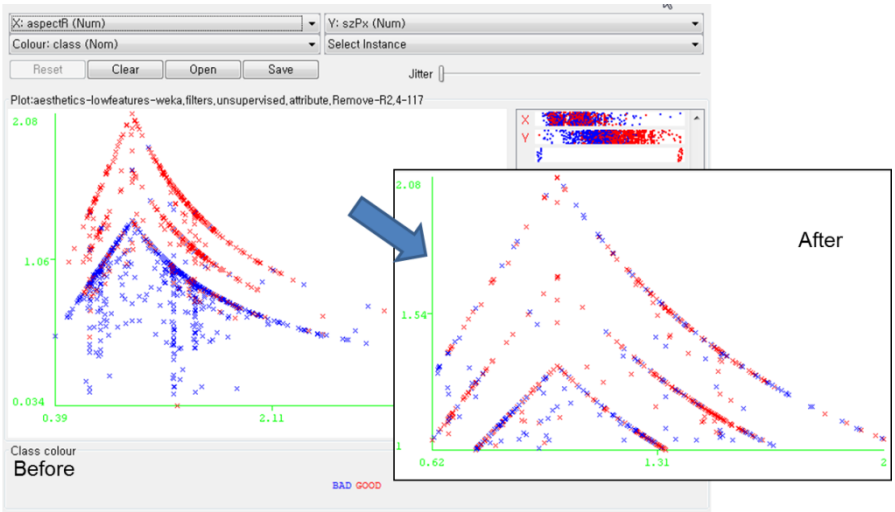


Figure 4. Visualized result of binary classification using two metadata: x-axis is for aspect ratio and y-axis for number of pixels. "Good" samples are colored in red while the "bad" in blue. The "After" pane is the result of controlling one of the parameter (number of pixels).

Lastly, it deserves to discuss the role of contextual metadata for aesthetic value estimation. For example, image dimensions including size and aspect ratio are reported as significant to estimate aesthetic value(Ciesielski et al., 2013). In the early stage of this research, aspect ratio and number of pixels of a photo were pointed out as two major features contributing to 78 percent of accuracy with the AVA dataset (good-vs-bad classification) as shown in the "before" pane of Figure 4. The peaks around the 1.0 of aspect ratio mean that the high resolution photos were provided in RAW format which is popular in DSLR cameras (and might be taken by professional photographers). Although this result implies the high correlation between quality features and perceived aesthetic value, this paper dismiss the metadata for

concentrating on finding content factors of determining beauty. The second pane named "After" in Figure 4 shows that discriminative power disappears if the number of pixels is controlled (normalized), exemplifying the limitation of approaches using contextual metadata for building visual aesthetics estimator.

CHAPTER 3. Data Revisited

3.1. What Makes Glass Ceiling

Among people who are trying to build a computational model for visual aesthetic perception, one of the unresolved issues is how to treat the mediocre samples, which are rated as 5 in a 10-point scale or 3 in a 5-point scale. Since Datta (Datta et al., 2006), researchers using machine learning techniques for computational aesthetics have consistently reported that excluding the mediocre group from the training set significantly helps enhancing the performance of the aesthetics estimator (Joshi et al., 2011; Murray, Marchesotti, Perronnin, et al., 2012). From the view of machine learning, the mediocre group causes two technical issues: imbalance problem and inappropriate sampling. The mediocre and the other groups are largely different in size, reaching 8 to 2 or even 9 to 1 in their ratio usually. Because such a huge imbalance causes significant deterioration when training classifiers with most machine learning techniques, rebalancing by spreading, subsampling, or upsampling like SMOTE (Chawla et al., 2002) are regarded as essential to preprocess the data (F. Provost, 2000). However, rebalancing usually raises another tough issue of selecting most representative samples appropriately: in other word, it is required to find a way of choose the topmost for the ambiguous mediocre samples for reducing the size of mediocre group while preserving the property. Due to the limitation, previous researches on computational aesthetics have excluded the mediocre group in their whole experiments (Datta et al., 2006) or at least in the training stage (Murray, Marchesotti, Perronnin, et al., 2012) while using it in the test stage.

In the context that almost all samples are rated as mediocre during aesthetic

evaluation, such exclusion might misguide the modeling by dismissing the fundamental issue in the real world of aesthetic appreciation. A binary aesthetic group classifier trained solely from the two extremes might suffer from frequent false positives due to the intrinsic majority of the mediocre in new incoming samples. Regression, an another alternative to classification approach (Datta & Wang, 2010; O. Wu, Hu, & Gao, 2011), is also affected by the issue. For example, ACQUINE (Datta & Wang, 2010), the online photo aesthetic analysis engine, regarded the distance from the hyper plane in the SVM classifier as an aesthetic score ranging from 0 to 100. However, in the following study, a large variance was observed in the scores for mediocre group and low correlation between ground truth and predicted scores was reported (Sachs, Kakarala, Castleman, & Rajan, 2011).

This difficult situation is calling for the study on the nature of aesthetic evaluation and its result captured by Likert scale survey, binary decision, or poll, performed in large scale owing to the popularity of social networks and photo services like Flickr. Especially, typicality of the good, the mediocre, and the bad should be validated because it determines learnability of the sample groups when training a computational model of aesthetic evaluation from the data. One of the methods of measuring typicality is measuring the consensus level among rater for each group. The consensus level of rating was previously mentioned by Ke and his colleagues (Ke et al., 2006); they dropped out the middle 80 percent of 60,000 photos in terms of average score and thereby made their aesthetics dataset more learnable from the view of machine learning, assuming that the mediocre group lacks consensus intrinsically. Their assumption regarding the average score as the consensus metric needs to be discussed further because such assumption requires another strong assumption that

score distributions for various factors are identical while disregarding the potential effect of information confidence, as an example. Another researchers (O. Wu et al., 2011) are joining the criticism by calling average score as an invalid measure due to the weakness and limitation they found.

Considering the huge amount of rated photos to be used for the typicality test, a new method is requested for measuring and visualizing consensus level of aesthetic scores among raters efficiently. For the purpose, I proposed a skewness-kurtosis map as the method, while rejecting more popular choices for measuring consensus like variance because it is vulnerable to bounded and skewed data (T. Park & Zhang, 2015). Instead of variance, kurtosis (the fourth moment, m4) is regarded as a good alternative because it indicates the lack of shoulders or infrequent extreme deviations (Balanda & MacGillivray, 1988): e.g., if almost all raters scored 5 points, its kurtosis would be significantly higher than the case that only a half of raters voted to the same point. The interpretation of kurtosis should consider its skewness (the third moment, m3) because there is a well-known relation of $K = S^2 + 1$ for Pearson's fourth moment used in this paper. Under the assumption of unimodality, skewness can be regarded as a representative property of showing a major score in a group, regardless of its asymmetry which distort mean (the first moment, m1). Leveraging the two properties altogether, the skewness-kurtosis map can be useful as a visualization tool for consensus analysis. The map has been used in other fields such as plasma physics, atmospheric science, oceanography, or financial engineering where deviations from Gaussianity should be investigated, while believing that this is the first case which adopted the S-K map to aesthetics: See (Cristelli, Zaccaria, & Pietronero, 2012; Sattin et al., 2009) for brief review.

Interquartile range (IQR), another popular dispersion measure, was initially tested as a consensus metric and then substituted by kurtosis, because it cannot differentiate two distributions with same (Q3 - Q1) and different ranges each other, which are not rare in aesthetic datasets. IQR's robustness to outliers seems deteriorating the representativeness as a shape descriptor of various distributions. Discrete score from 1 to 10 in Likert-scale is also responsible for reducing the representation power of the rank-based consensus measure.

3.2. Consensus Analysis

3.2.1 Data Set

Among several available massive visual aesthetics datasets, AVA (Aesthetic Visual Analysis) dataset (Murray, Marchesotti, & Perronnin, 2012), which has been publicly available since 2012, was selected. It consists of 255,530 photos and their 10-point (1 to 10) scores of aesthetics, rated by 200 (in average) photography professionals and hobbyists via online during a certain period of "challenge" in the website www.dpchallenge.com. A photo was displayed in the web page with grey padding while preserving its original resolution and aspect ratio during the rating. A new rater was not exposed by the accumulated result.

Contrary to Photo.net (Datta et al., 2006) or CUHK dataset (Ke et al., 2006), the AVA dataset preserves all score distributions, which is essential for consensus analysis, making the dataset privileged. It also provides 65 textual tags (e.g., landscape, street,

portraiture, food) describing the subjects or styles explicitly and thereby helps finding semantic factors in addition to consensus analysis. Approximately 8,000 images in average are provided for each tag, and the mean distributions of aesthetic scores for tags are balanced: See Murray et al. (Murray, Marchesotti, Perronnin, et al., 2012) for the details.

For minimizing bias to a specific dataset, additional large-scale aesthetics datasets have been searched as more generalized source of evidence, concluding that AVA dataset is currently the only available massive aesthetics dataset from the view of scale and quality, unfortunately. The providers of AVA dataset summarized (Murray, Marchesotti, Perronnin, et al., 2012) explain that other previous datasets lack in scale or suffer from intrinsic bias which may misguide consensus analysis. For example, Photo.Net dataset (Datta et al., 2006) suffers from small size of raters (50 in average), artificial frame embedding for several photos which cause bias in aesthetic judgment (Marchesotti et al., 2011), and positive correlation between average scores and number of ratings (Joshi et al., 2011); CUHK dataset (Ke et al., 2006) intentionally exclude mediocre photos, the essential part of the proposed consensus analysis; and, ImageCLEF (Müller, Clough, Deselaers, Caputo, & CLEF, 2010), a Flickr-based large-scale dataset for concept detection, captures more abstract "interestingness" instead.

3.2.2 Method

The excess kurtosis (the Pearson measure), called as "kurtosis" through this paper, is used for easy visualization because all observed samples are leptokurtic in the map. For all 255,530 photos in AVA dataset, all four moments and quantile of score distribution for a photo are calculated and plotted in the S-K plane; median and textual tags were additionally used to see group tendency.

Three reference trajectories in the S-K map were used to visualize the characteristics of the score distributions in comparative manner; Gaussian trajectory, Klaassen bound (Klaassen, Mokveld, & Van Es, 2000), and 4/3 power law trajectory. Gaussian parabolic relation $K=S^2$ for unbounded Gaussian random samples. Klaassen bound is a theoretical lower bound of any unimodal distribution in S-K plane which is generated by the function (Klaassen et al., 2000)

$$K= S^2 + 189/125 \quad (8)$$

for the purpose of unimodality check in the S-K map, the offset term of "-186/125" in the original equation is adjusted to "189/125" as a lower bound because excess kurtosis is used for K. This bound is usually drawn as a green solid line for all figures of S-K map in this paper.

The 4/3 power law trajectory is generated from the function (Cristelli et al., 2012) of $K = 60^{1/3} S^{4/3}$. The last 4/3 power-law regime, drawn as the red dashed line in this

paper, was originally reported in the recent article (Cristelli et al., 2012) about the financial market data: the daily returns of S&P 500 stocks. Contrary to most physical data showing parabolic near Gaussian region, a few data from earthquake record and the stock market are known to converge to the power law trajectory in S-K plane.

Considering the ubiquity of power-law relation in experimental psychology, it deserves to see whether or not the aesthetic evaluation, as one of various mental activities, shows the pattern which might signify a dynamic process behind the phenomenon.

Lastly, Kruskal-Wallis test is selected for the case where hypothesis test is required. Kruskal-Wallis test is one of the non-parametric statistical methods for comparing two or more samples by rank. More popular counterparts in parametric methods such as t-Test or one-way analysis of variance (ANOVA) are hardly applicable to the AVA dataset analysis because they assume a normal distribution of the residuals and therefore mislead to false interpretation when applied to the case of non-Gaussianity.

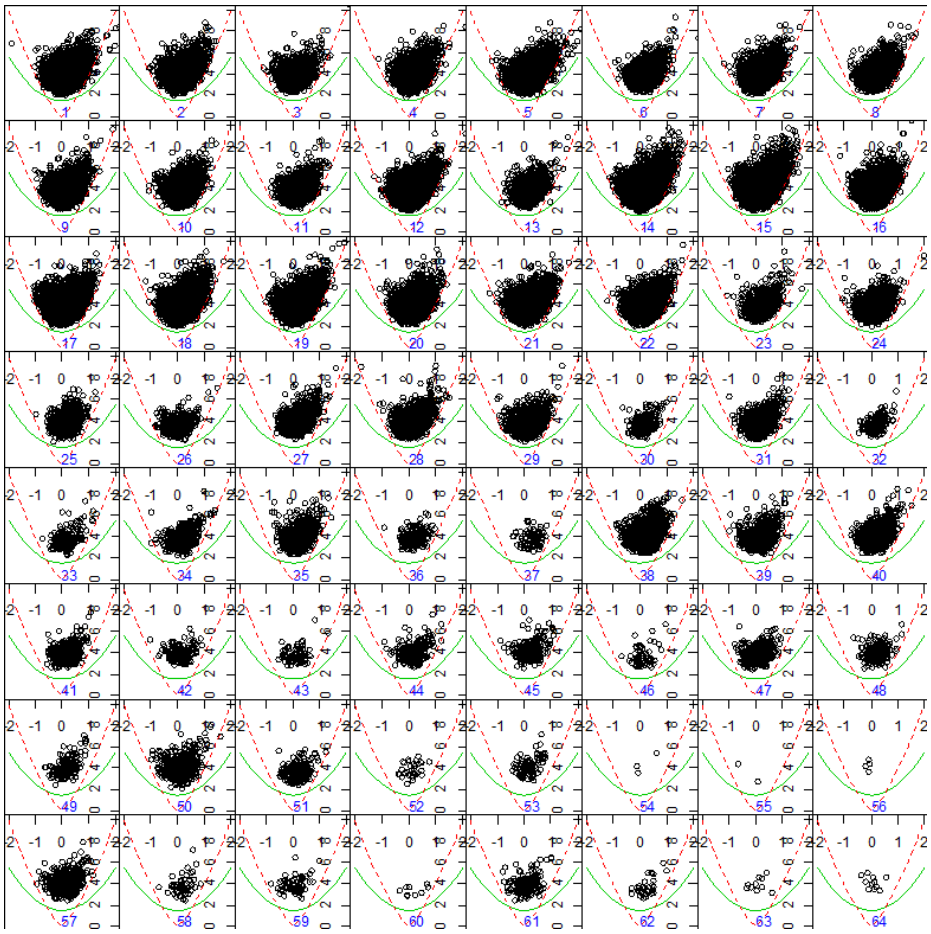


Figure 5. The S-K maps of aesthetic score distribution for 64 tag groups

3.3. Analysis Results: 4 Patterns

Figure 5 shows the S-K maps for all 64 tags (the last tag 65 is excluded due to insufficient samples) which share the same axis of [-2, 2] in m3 (skewness, the 3rd moment) and [0, 10] in m4 (kurtosis, the 4th moment). Due to the characteristics of S-K plane, the "good" photos locate in the negative m3 pane of the S-K map while the "bad" in the positive, which is inverse to the more familiar direction in raw score

distribution.

Throughout the analysis of aesthetic score distributions from AVA dataset, four patterns are observed when projected to the S-K plane as following:

3.3.1 Pattern 1: A Wide Kurtosis Range

The most notable pattern in Figure 5 is the wide kurtosis range, from 2.0 to 10.0 approximately; even it reaches up to 8 for the symmetric samples (m3 is near zero). Considering the nature of kurtosis as a consensus metric, it is hard to expect such wide range in the mediocre group because it means there is a universally agreed mediocrity, while other mediocre samples are not making consensus among raters.

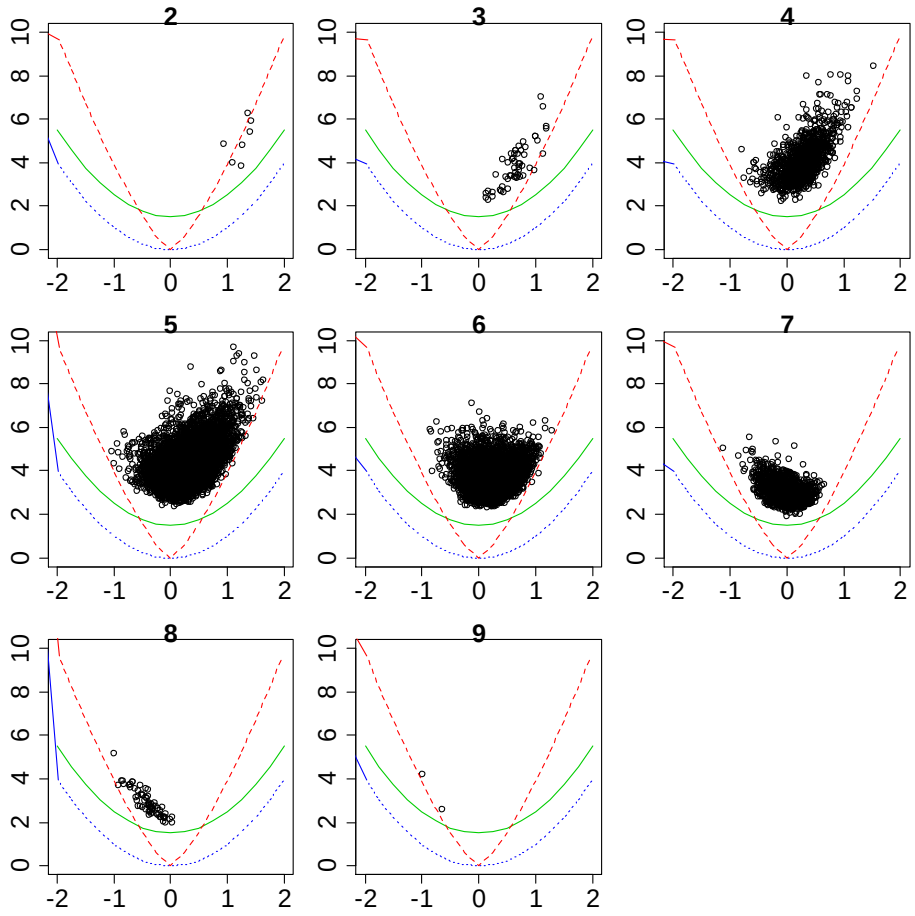


Figure 6. Skewness-kurtosis relation of aesthetic score distributions for eight "landscape" photo groups clustered by median score from 2 to 9 (top of each plot): x-axis for m3 (skewness) and y-axis for m4 (kurtosis). Red dashed lines are power law trajectories, green solid lines are Klaassen bounds, and blue dotted lines are Gaussian parabolic relations.

This clear non-Gaussianity is universal across all the tags to various degrees. Figure 6 depicts the relation between skewness (m3) and kurtosis (m4) for the photos with the tag 14 ("landscape") as one of the most representative tag groups. For visual

clarity, they are clustered by median score (on top of each box) and the Gaussian parabolic relation $K=S^2$ (the bottommost blue dot line) is added to the other guidelines inherited from Fig. 1.

The bias of neutral point to the 6 median group is regarded as the side effect of the 10-point Likert scale questionnaire that DPChallenge.com adopted because of the arithmetic middle point is not 5 but 5.5 in this scale . The proportionality between m3 and m4 in both extreme groups (2~3 and 8~9 in their median score) is a natural result of the mathematical relation between the two variables unless its various scales specific to the tags is considered.

Figure 6 reveals that the wide range of kurtosis along the axis of m3 = 0 in Figure 5 is mainly caused by the mediocre group, not by the exceptional combination of the good and the bad groups. It means that the degree of consensus in the mediocre group varies greatly; in other word, the mediocrity originates from not only the "lack of consensus" but also the (almost) unanimous agreement.

For validating the interpretation on the wide kurtosis range, two contrastive samples are selected for comparing the raw score distributions and their normality. Figure 7 shows the score distributions and the Q-Q plots with normal distribution for two contrastive samples selected from the lowest (m4 = 2.44) and the highest (m4 = 6.68) kurtosis region respectively along the virtual axis of m3 = 0 in the S-K map.

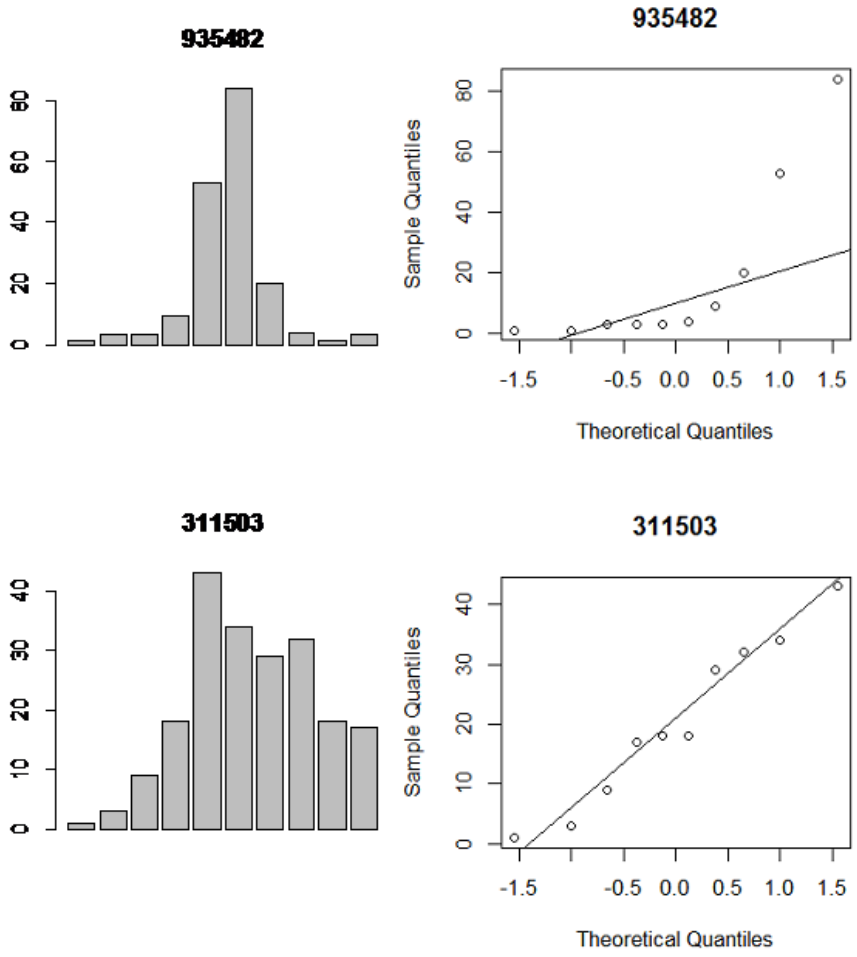


Figure 7. Score distributions and Q-Q plots of normality test for two contrastive samples

In Figure 7, the upper row represents the sample (ID: 935482, number of raters = 181) at the point of m3 = -0.02 and m4 = 6.68 in the S-K map, and the lower row does the opposite sample (ID: 311503, number of raters = 252) at m3 = -0.01 and m4 =

2.44. As clearly depicted in the Q-Q plots between the observed distribution and the fitted normal distribution, the high m4 sample failed in Shapiro-Wilcox normality test by p-value = 0.0004 while the low one passed it by p-value = 0.7143 (all with 95 percent confidence intervals). Although all photos are excluded from this paper due to a potential copyright issue, the photos are freely accessible at the website www.dpchallenge.com: for an instance, the photo 935482 is accessible by using the URL www.dpchallenge.com/image.php?IMAGE_ID=935482.

3.3.2 Pattern 2: Consensus Asymmetry

Another property observed in Figure 6 is the asymmetry of kurtosis range between the two contrastive groups; the m4 of a "bad" sample tends to be higher than that of a "good" one. To measure the degree of asymmetry for each tag, the kurtosis distribution of two extreme groups are visualized by boxplot pairs of two extremes as depicted in Fig. 4: for reasons of space, only 33 tags are plotted.

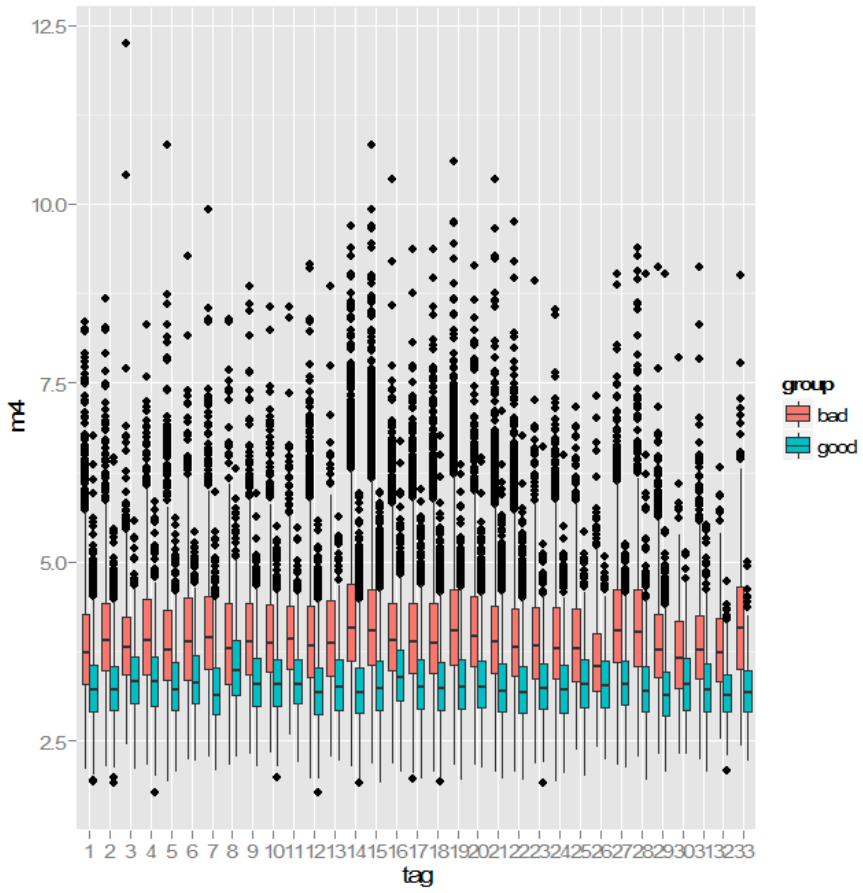


Figure 8. Boxplot pairs of the m4 distribution for 33 tags. Each pair consists of the bad group (the left red box) and the good one (the right blue box).

To ignore the tag-specific preferential bias, two quartiles of mean score for each tag group are used as thresholds: m1 > Q_3 (75th percentile) for the good group and m1 < Q_1 (25th percentile) for the bad. Figure 8 confirms that a score distribution for a bad photo tends to have significantly higher kurtosis than a good one for all tags except several style tags: in Kruskal-Wallis test between the two groups for all tags, only the

tag 54 (texture library), 55 (overlay), 60 (pinhole), and 62 (lensbaby) failed rejecting the null hypothesis with a 95 percent confidence interval. It can be interpreted as superiority of negative aesthetic evaluation to positive one, or vice versa, from the view of making consensus.

3.3.3 Pattern 3: The 4/3 Power Law Regime

Figure 6 shows that, for "landscape" photos in AVA dataset, there is a positive correlation between the proximity of samples to the 4/3 power law trajectory in the S-K map and the sore offset from the "neutral" point (5.5 or 6 in 10-point Likert scale).

For example, the most samples from the group of median score is 3 or 8 locate near the power law trajectories while samples from the neutral score group (median score is 6) are relatively scattered in the S-K map. Even though such a convergence to the power law trajectories (red dashed lines) in both extreme groups ("very good" or "very bad") can be partly explained as a truncated normal distribution which cause ceiling and flooring at the sore of 1 and 10 respectively, a significant number of samples near the trajectories from the neutral score group raise an issue of strong non-Gaussian property behind it.

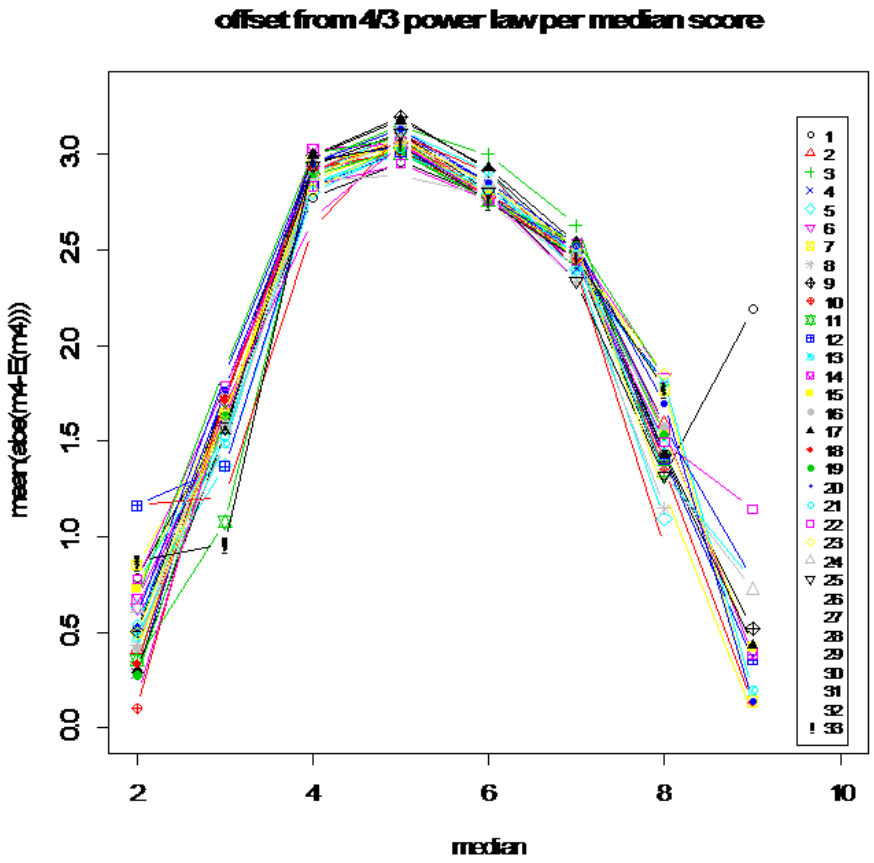


Figure 9. Mean offset from the 4/3 power law per median score

Figure 9 illustrates the common tendency of the convergence across the first 33 tags by comparing the mean distances between observed kurtosis and the estimated value from the 4/3 the power law per median score. It also shows that inter-tag difference significantly increases for both extreme score groups.

Additionally, it is evident all samples locate above the unimodal distribution bound (Klaassen et al., 2000) of $K = S^2 + 189/125$ in Figs. 1 and 2.

3.3.4 Pattern 4: Tag Effect

While the above three patterns seems almost universal among the photos, the degrees of the patterns are affected by tags. For consensus, the Kruskal-Wallis rank sum test on the effect of tags to kurtosis, the consensus metric in this paper, rejected the null hypothesis of equality by p-value = 2.2e-16 (with a 95 percent confidence interval).

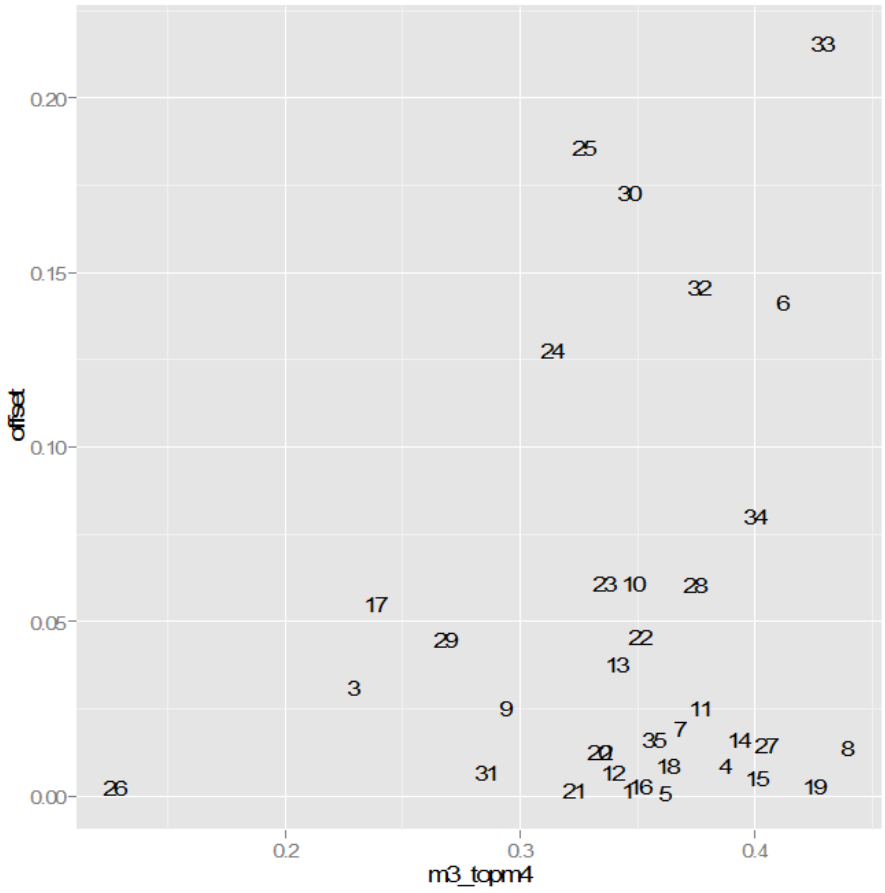


Figure 10. Scatter plot of asymmetry and 4/3 power law offset

Figure 10 show each tag group in the scatter plot of the position of the offset from 4/3 power law and "m3_topm4," the mean skewness of the samples in the top 25 percentile kurtosis.

Therefore, a sample locates in the lower right pane of the plot if it is close to the power law regime and asymmetric. The distribution of tags in the plot shows a tendency that more asymmetric and power-law-compliant photo groups, usually located in the lower right quadrant, tend to be assigned with the tags about natural and spatial attributes including snapshot ("8"), animals ("19"), landscape ("14"), nature ("15"), rural ("27"), sky ("7"), still life ("18"), and seascape ("35"), while the most symmetric or far-from-power-law group at the lower left or upper right quadrants tends to be followed by the tags about abstract concepts including action ("24"), photojournalism ("25"), political ("30"), sports ("9"), and fashion ("3"). The mean amount of images for natural and abstract groups are 2287 and 518 respectively.

3.4 Discussion

Analysis in the higher moments, skewness and kurtosis, provides a new perspective to the aesthetics data interpretation. While Murray et al. (Murray, Marchesotti, Perronnin, et al., 2012) saw the aesthetic score distribution for a photo in the same AVA

dataset as “largely Gaussian” after the review in the low moments regarding strong non-Gaussian characteristics in the both extremes as the floor-ceiling effect, a strong non-Gaussianity (very likely to be a member of a power law family) was found in the same data showing a mixture of non-Gaussian and Gaussian distributions in the S-K plane. They also reported another important but misinterpreted property that the standard deviation of score distribution is proportional to the offset from the mean. The innate distortion of variance for highly skewed data and loss of information about “variance of variance” in the low moments are thought to be responsible for such insufficiency in the previous explanation.

Especially, kurtosis of score distribution is thought to be a good measure of consensus among raters and should be treated as a key factor in modeling aesthetic evaluation process. The presence of a the new factor can explain why learnability of CUHK dataset (Ke et al., 2006) is so high, by interpreting it as the result of excluding all low consensus samples and thereby resampling tailored data only. In addition, the conceptual similarity between the concept of consensus and the confidence in signal detection theory deserves to be investigated further.

Another major pattern of asymmetry in kurtosis toward negative evaluation is in coordination with the lesson acquired in other researches on emotion. In the field of human computer interaction (HCI), there is a similar report that negative aesthetic

decision on webpage design is made faster than the positive one (Tractinsky, Cokhavi, Kirschenbaum, & Sharfi, 2006). Cerosaletti and Loui also reports the similar phenomenon in terms of an "inverted U-curve" in standard deviation of rating along score range (Cerosaletti & Loui, 2009).

For the third pattern, the regime of the 4/3 power law, Cristelli et al. (Cristelli et al., 2012) insist that it implies the presence of interaction between multiple agents behind the phenomenon. Agreeing with the opinion, it seems that visual aesthetic evaluation can be modeled in the similar manner, while treating the consensus issue separately.

For the fourth pattern of tag effect, the discriminative relation between natural objects and abstract concepts is in accordance with the previous studies which tried to explain the inborn preference in the framework of evolutionary psychology (Tooby & Cosmides, 2001; Vessel & Rubin, 2010): especially, an innate preferential bias toward landscape has been studied thoroughly as the result of adaptation. Art appreciation seems more subjective to individuals: it has been traditional consensus that taste can be developed by training (Eysenck, 1940) and that experienced individuals and art-naive individuals are clearly different in art appreciation (Hekkert & Van Wieringen, 1996).

Another qualitative report mentioned a similar but unclear tendency that a photo with more variance in its ratings is usually non-conventional (Murray, Marchesotti, Perronnin, et al., 2012). Such an interpersonal similarity implies that consensus

might be the matter of more "hardwired" attractors at least for some subjects of appreciation.

Lastly, to validate the score of AVA dataset as aesthetic quality measurement, the effect of theme relevance to aesthetic rating was analyzed by comparing free and non-free studies: If most images in AVA dataset were rated not only in their aesthetic quality but also in the relevance of the image to a given theme, this may raise a question whether or not the dataset really capture aesthetic visual perception. The challenge types of all 255,530 photos in AVA dataset have been searched and classified them into two groups: the free study and the non-free. From the view of group size, the result shows that the ratio of the free study group to the non-free is 29,351: 226,179 (1:7.7 approximately). As visualized in the two S-K maps from the two groups in Appendix A, any significant difference between them was not found as far as the samples in a subset are sufficient: for both groups, several tags lack samples enough to show the patterns as observed in the other tags. For this issue, AVA providers (Murray, Marchesotti, Perronnin, et al., 2012) analyzed the effect of relevance to a theme as following:

> *"While we observed no trend among challenges with high-variance score distributions, we found that the majority of free study challenges were among the bottom 100 challenges by variance, with 11 free studies among the bottom 20 challenges. Free study challenges have no*

restrictions or requirements as to the subject matter of the submitted photographs. The low variance of these types of 2412 challenges indicates that challenges with specific requirements tend to lead to a greater variance of opinion, probably with respect to how well entries adhere to these requirements.''

In other words, free studies (no relevance) are reported to show relatively higher consensus; considering relevance seems lowering consensus among raters. Therefore, relevance effect, if exist, seems not negating the patterns (of wide kurtosis range, at least): See Appendix A for the S-K maps from the two groups.

CHAPTER 4. Modeling

4.1. Background

For comprehensive explanation on all the patterns reviewed in the previous chapter, several prior arts of modeling visual aesthetic evaluation were investigated, finding that just a few researchers (Leder et al., 2004; Pelowski & Akiba, 2011; Reber et al., 2004; Winkielman, Halberstadt, Fazendeiro, & Catty, 2006) have proposed conceptual models of visual aesthetic evaluation. For example, Leder et al. (Leder et al., 2004) proposed an information-staging process model of art perception consisting of the five-stage modules in cascaded manner. Pelowski and Akiba (Pelowski & Akiba, 2011) proposed a similar multi-stage model in the same domain. Reber and his colleagues (Reber et al., 2004) insists that fluency in information processing determines attractiveness of stimuli, followed by Winkielman et al. (Winkielman et al., 2006) who suggests innate attractiveness of prototypes because they can be processed fluently with less workload. However, unfortunately, there wasn't any previous aesthetic evaluation model which is able to explain the characteristic patterns observed in the massive dataset quantitatively. While Wu et al. (O. Wu et al., 2011) share the same intuition that there is a sample-specific difference in consensus level, their genuine approach of multi-label classification regards these as given, not trying to model the underlying mechanism.

Therefore, several models were devised for the purpose of acquiring a quantitative alternative. Motivated by the 4/3 power law regime and the exceptionally wide kurtosis range, the proposed approach for the modelling is based on the dynamic process of multiplicative interaction between several positive and negative attractors with ambient Gaussian noise. In this scheme, the level of consensus and inter-tag difference in the S-K plane is determined by the tag-specific configuration of attractors. It is expected that a successful model should be able to simulate the four patterns observed in the AVA dataset.

For understanding the relationship between models and their characteristics from the view of the four-pattern representation, a static model with multiple attractors is analyzed firstly and then it expands to a dynamic model.

4.2. Static Models

The simplest model would assume one aesthetic factor which varies in its degree of effect following a certain probabilistic distribution. For an instance, absolute difference between an imaginary ideal spatial layout and an observed one in an image, if exist, can be regarded as a single static factor of determining perceived aesthetic value, followed by random noise.

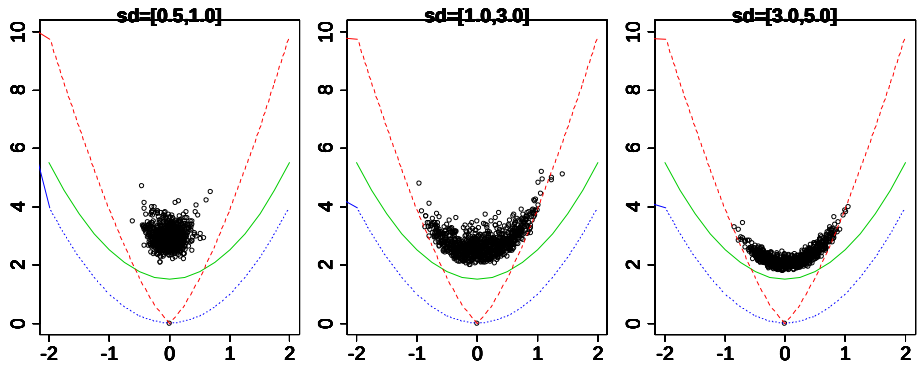


Figure 11. S-K plots for Gaussian distributions with three different standard deviations.

Figure 11 shows the S-K plane projection of Monte-Carlo simulation results using the single-factor static model using Gaussian distribution for the mediocre group with three different standard deviation ranges, the very indicator of consensus in Gaussian distribution, while sharing a same mean of 5. In this simulation, three different ranges of standard deviation with eleven steps are used for Gaussian random number generation and number of trials for each configuration is determined as 200, following the average number of voters for each photo in AVA dataset. Considering the bounded nature of scores between 1 and 10 in the voting system used in AVA dataset, truncation in the same range of 1 and 10 as minimum and maximum is applied to the second (sd = [1.0 3.0]) and the third (sd = [3.0 5.0]) configuration; without truncation, increasing variance does not affect the projected form in the S-K plane by showing as

same patterns as the first configuration (sd = [0.5 1.0]) in Fig. 7.

Their scattered pattern in the S-K plane reveals that changing standard deviation in Gaussian distribution does not help simulating the wide kurtosis range and asymmetry observed so far, although it simulates the 4/3 power law regime, mainly due to the truncation effect.

Considering the accumulated evidences about the plurality of aesthetic factors in the field of experimental psychology, a multiple factor model rather than the above single factor model is regarded as more appropriate to explain human visual perception of aesthetics. In case of color and brightness, since Eysenck (Eysenck, 1940), many researchers (Guilford & Smith, 1959; Hurlbert & Ling, 2007; McManusU et al., 1981; Ou et al., 2004; Palmer & Schloss, 2010) have reported that there is a systematic pattern in group color preference: in hue preference, green, cyan, and blue are usually preferred to red and yellow (Hurlbert & Ling, 2007; Palmer & Schloss, 2010); saturated colors are generally preferred (McManusU et al., 1981; Ou et al., 2004); and, hue-value interaction exists in the way that a brighter image is generally preferred with different peak points for each color (Guilford & Smith, 1959; Palmer & Schloss, 2010). In case of spatial structure, preference to horizontal and vertical lines (Latto, Brain, & Kelly, 2000), 1/f power spectra preference (Fernandez & Wilkins, 2008; Graham & Field, 2008), golden ratio (Atalay, 2004; Konečni & Cline, 2001), symmetry(Jacobsen

& Hofel, 2002; Palmer & Griscom, 2013), soft curvature (Bar & Neta, 2006; Leder et al., 2011; Silvia & Barona, 2009) and canonical composition (Bertamini, Bennett, & Bode, 2011; Konkle & Oliva, 2011; Linsen et al., 2010; Sammartino & Palmer, 2012) have been reported as significant factors for visual aesthetic perception.

Such a multiple factor model can be implemented in various approaches. For an instance, the net value of perceived aesthetics can be a weighted sum of two or more factors; e.g., a spatial layout and a color tone. The proposed approach is clustering multiple factors as the two groups, positive and negative, and regarding these as two group factors. Among the several probabilistic distributions which treat two or more factors, beta distribution is selected because it is modeled by the product of two contrastive bases, X and (1-X), which is in accordance with the contrast between the positive and the negative factors, while allowing different powers for the two base factors which are convenient to simulate the observed asymmetry. The probability density function (pdf) of beta distribution is a power function:

$$f(x;\alpha,\beta)=Cx^{(\alpha-1)}(1-x)^{(\beta-1)} \quad (9)$$

where the normalization constant C is the product of three gamma functions $\Gamma(\alpha+\beta)$, $\Gamma(\alpha)^{-1}$, and $\Gamma(\beta)^{-1}$. From the view of aesthetics modeling, the simulated perception of visual aesthetics can be interpreted as the product of "good"(x) and "bad" (1-x) with their own powers. In this model, the two shape parameters, α and β, are regarded as the

degree of effect that the two factor groups evoke in their respective directions; e.g., strength and numbers of factors in a positive group determine the shape parameter α collectively.

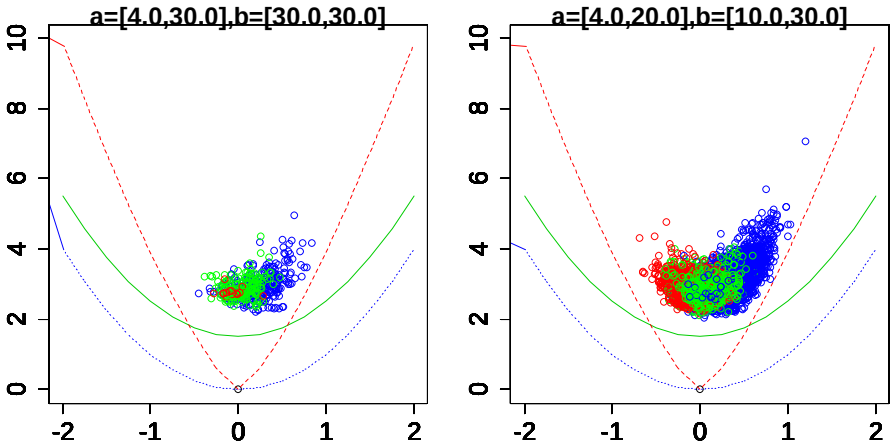


Figure 12. S-K plots for two combinations of the alpha and the beta ranges. Colored by their median aesthetic scores v: red for v≥6, blue for v≤4, and green for the others.

Figure 12 is the simulation result of beta distributions generated from two different configurations. The left configuration in Figure 12 changes the power of "good"(α) while keeping the power of "bad" (β) as constant. The right one changes both powers with structural bias to the dislike factors by higher maximum value of β. In comparison with Figure 11, Figure 12 shows a better result in that it meets all requirements except the wide K range in the mediocre group; especially, the second pattern, consensus asymmetry, is easily represented by rebalancing the power ranges of the two factor

groups.

Although the simple beta distribution model was used with two factors for clarity of explanation, if one of the two factors consists of multiple subfactors, this model can be easily extended to Dirichlet distribution, the multivariate generalized version of beta distribution.

4.3. Dynamic Models (DDM4AP)

The last unresolved issue of the wide kurtosis range in the static models for aesthetic evaluation reveals the need for another computational model which is able to produce significantly different evaluation results in accordance with small variance in model parameters among people. Also, it seems desirable to inherit the concept of two contrastive multiple factor groups as shown in the previous beta distribution model because it is useful for simulating consensus asymmetry.

At this point, dynamic systems are considered as the candidate for revising the model based on knowledge that it is easy to observe such a parameter-sensitive change (e.g., bifurcation) with the models of dynamic systems (Kelso, 1997). Among various dynamic models, a Drift-Diffusion Model (DDM) (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratcliff, 1978; Ratcliff & McKoon, 2008) has been popular among psychologists as a well-defined model of explaining behavioral data for the task

of forced categorization among two or more alternatives (See Ditterich (Ditterich, 2010) or Bogacz et al. (Bogacz et al., 2006) for review). This model assumes that human mind requiring a binary decision (A or B) accumulates evidence favoring each alternative over time, while simultaneously distracted by internal random walk (noise), and makes decision when the accumulated evidence for either alternative is enough. This process is usually depicted as a particle drifts and diffusions between the two boundaries until it reaches either boundary. In this view, each boundary attracts the particle at a given or varying rate with noisy turbulence, making the process stochastic. For an instance, in a traditional DDM assuming only one attractor in one side, the evidence is accumulated in according to

$$dx = A\,dt + W, \quad x(0) = 0. \tag{10}$$

In Equation 10, x grows at rate A in average while white noise (W) is continuously added (Bogacz et al., 2006).

If aesthetic perception is regarded as a compromise between "good" and "bad" factors which attract in opposite directions, the DDM can be used to model the dynamic process with the variance in the number and positions of the attractors in both sides. Another aspect to be considered during the choice of models is compliance with

neuroscientific evidence. A recent fMRI study of T-shirts appreciation implies that the model for human visual aesthetic evaluation might have to be different from the model for semantic understanding, because they report that the value of the visual aesthetic attributes and that of the semantic attributes do not share the same neural correlates (Lim, O'Doherty, & Rangel, 2013): specifically, the fusiform gyrus correlated not with the semantic attributes but with the visual aesthetic attributes, while the posterior superior temporal gyrus exhibits the opposite pattern.

Relatively, temporal properties of affect models have not been fully explored except a few models like WASABI (Becker-Asano & Wachsmuth, 2008) or componential appraisal theory (Scherer, 2005). For modeling the interaction between multiple attractors with moderate randomness, motivated from the diffusion decision model (Ratcliff & McKoon, 2008), I propose a new model named Drift Diffusion Model for Aesthetic Perception (DDM4AP). Because aesthetic perception is usually captured as a multi-label classification problem using n-point Likert scale (e.g., 10-point scale in AVA dataset), it is inevitable for the new model to modify the original drift-diffusion model, which assume a 2AFC (Two Alternative Forced Choice) task, significantly except its core concepts, although aesthetic perception can also be approximated to the binary decision of 2AFC: like or dislike.

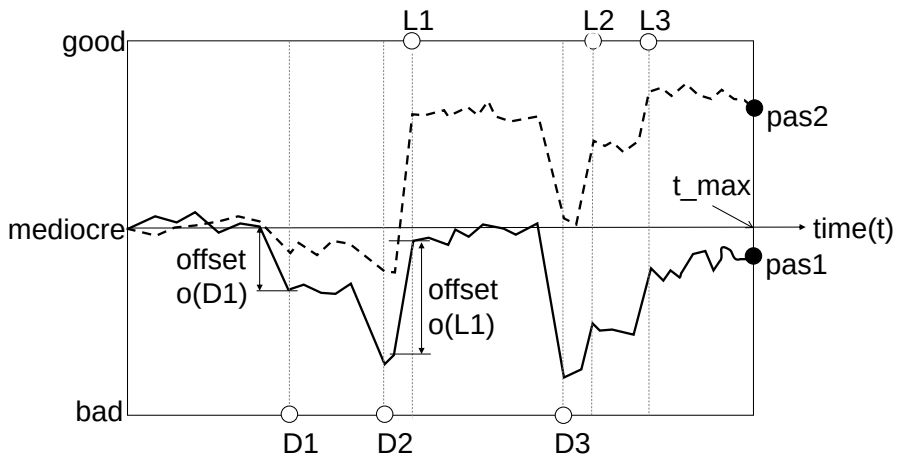


Figure 13. An example of DDM4AP with three "like" factors at the good side (top) and "dislike" factors at the bad side (bottom) respectively.

Figure 13 shows an illustrative example of DDM4AP having equal numbers of attractors in the both sides, the good and the bad. The perception process ends when the reference point (the black dot in Figure 13) reaches the end of time axis, t_max, or either border (top or bottom) before t_max. The finally perceived aesthetic value (v) is mapped from 1 to 10 along the three borders starting from the bottom ("bad") through the rightmost border to the top ("good"); as the 10-point score follows a Likert scale, the score distribution along the borders can be arbitrary as far as it preserves the order. If there are three L attractors and one D attractor, the landing position of the reference point will be systematically biased to be somewhere on the top or rightmost border while the offset is determined by the drift rate; if there isn't any attractor in the stimuli,

the perceived value of DDM4AP is solely determined by diffusion, which is usually modeled as bounded white noise.

In DDM4AP, two factor groups, L (Like) at the top side and D (Dislike) at the bottom, are treated as the collection of attractors: L_i for good and D_j for bad respectively. The number and the sequence of attractors along time axis at both sides significantly affect the resultant aesthetic value at the end of the process. In Figure 13, offset $o(D_1)$ depends on the probabilistic gain of the first attractor D_1 while offset $o(L_1)$ is determined by L_1 located on the opposite side.

Therefore, at the end of the process, the perceived aesthetic value v is determined as following:

$$v = \text{neutral score} + \sum_{i=1}^{m} o(L_i) - \sum_{j=1}^{n} o(D_j) + W \quad (11)$$

where m and n are the number of attractors in the good and the bad side respectively which exist before the end position at time axis, W as a uniformly distributed net diffusion, and $o(L_i)$ or $o(D_i)$ are random variables following Gaussian or exponential probability density function (pdf). The probabilistic drift rate, which is different from the constant drift rate of traditional DDM, assumes that the drift rate follows power distribution as the result of interaction among multiple agents (i.e.,

another small neural network of detecting a factor in a visual stimulus), or Gaussian distribution as the representation of mean drift rate with variance which is commonplace in nature. Combined with another change of impulsive accumulation of the drift offset at the position of an attractor, the characteristics of the upper and lower bounds are differentiated from that of the traditional DDM; from consistent factors to confidence bounds. For avoiding the confusion about different concepts of drift rates between DDM and DDM4AP, a new term "stochastic attraction rate" will be used for the probabilistic drift rate hereinafter.

Also, the number and positions of attractors at both sides are regarded as being determined by visual stimuli, while the drift rate and/or diffusion rate are personal traits. For example, in the simulation of DDM4AP, the number of attractors and their positions are determined for each photo, while the drift and diffusion rates are set for each trial (of a virtual human participant).

The neutral score in Equation 3 is usually set to 5 or 5.5 in the 10-point scoring system.

Due to the characteristics of drift-diffusion system, the temporal distribution of attractors is also an important factor affecting not only latency but also the offset from neutral evaluation. For example, the perceived value (v) in Figure 13 would be systematically biased to "bad" if the D group locates prior to the L group. The idea of

temporal distribution of attractors is justified by the various processing times in all levels of perception and cognition.

Because the position and gain of the attractors are assumed to be different photo by photo, the simulation results are generated by overlapping the responses from 200 people for 100 photos.

In the Monte-Carlo simulation, the positions of attractors are determined by uniform random number generation between 100 and 300 trials while their gains follow the exponential distribution with $\lambda = 5.0$ or Gaussian distribution. The noise term is implemented as uniform random distribution (white noise) with the gain of 0.015.

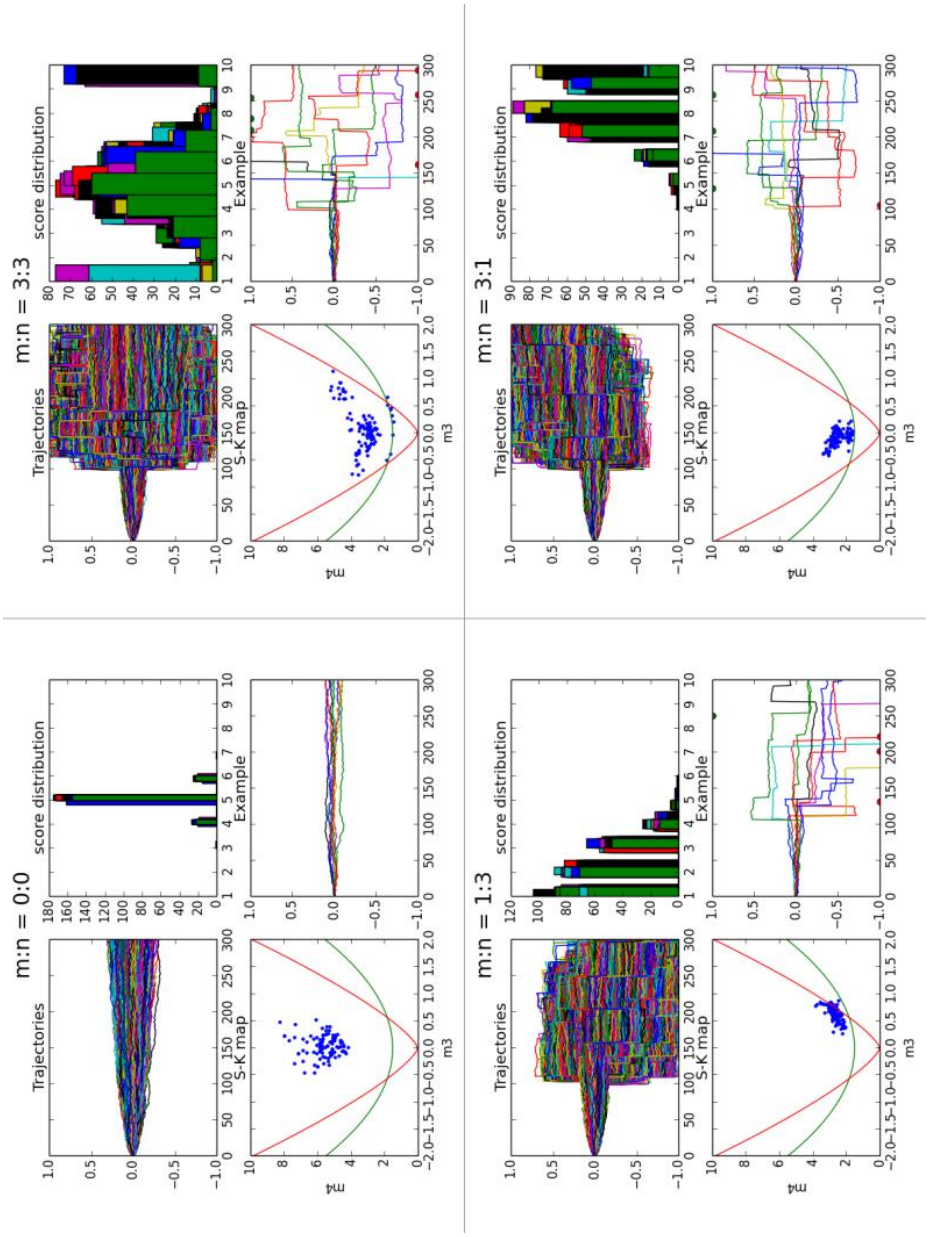


Figure 14. Simulation results of DDM4AP with Gaussian attractors.

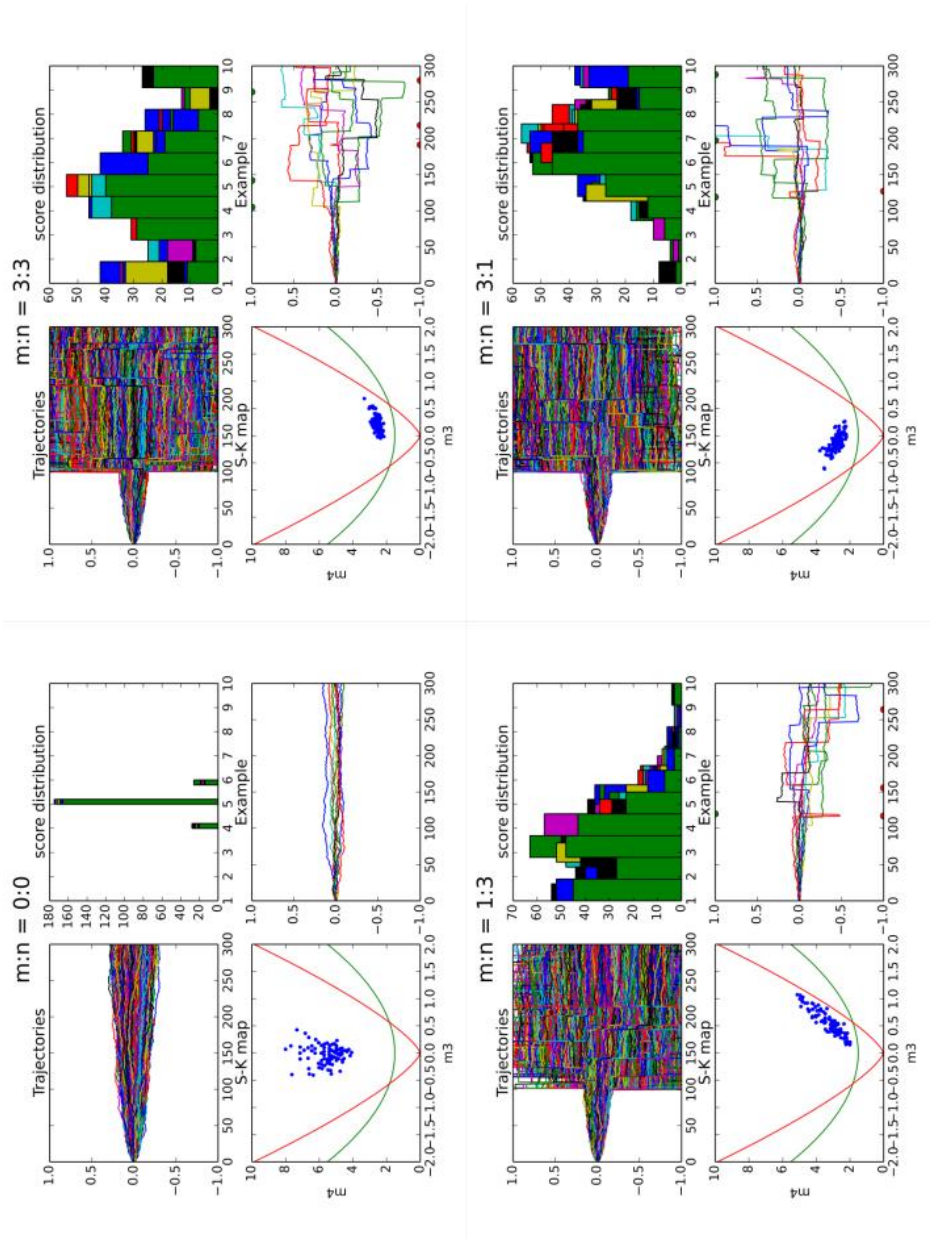


Figure 15. Simulation results of DDM4AP with exponential attractors

Figure 14 and Figure 15 show the simulation results of DDM4AP with two different types of attractors (Gaussian and exponential) which meet all requirements as proposed. For each m:n case, the fourth quadrant shows a simplified DDM4AP trajectories down to ten raters with the position marks of L_i (the green semi-circles on the upper bound) and D_i (the red semi-circles on the lower bound). Contrary to the static model, it simulates the wide K range for the mediocre group (L and D are balanced or equally void) in the S-K map while mimicking asymmetry by rebalancing drift between the two attractor groups. Comparing the two figures, assuming exponential attractors results in more realistic score distributions rather than its Gaussian counterpart; contrary to the exponential version in Figure 15, the Gaussian version in Figure 14 produces too biased score distributions in the setting.

At least in the same setting, a specific case of 0:0 in the ratio of attractor numbers between the two groups is responsible for such a high kurtosis in the mediocre group, given that the diffusion rate is small enough compared with drift rates. This result is interpreted as natural in DDM4AP because the other balanced case (m = n) might generate more various results due to it is apt to be affected by small difference in stochastic attraction rate and process time (therefore the position in time axis) for a common factor among people.

For measuring explanation power of the dynamic models quantitatively, L2-norm

error was defined as following:

$$error = \sqrt{(k_{min} - t_{min})^2 + (k_{med} - t_{med})^2 + (k_{max} - t_{max})^2} \quad (12)$$

where k_{min}, k_{med}, and k_{max} are minimum, median, and maximum of kurtosis of simulated score distributions respectively, while t_{min}, t_{med}, and t_{max} are their pairwise counterparts calculated from AVA dataset. For controlling the effect of tag and asymmetry in the number of attractors, 1679 "landscape" photos with the median score of five were selected to calculate kurtoses of their score distributions, resulting in t_{min} = 2.43 , t_{med} = 4.03, and t_{max} =8.75.

For finding optimal parameters of two independent variables, stochastic attraction rate and diffusion rate, grid search was used for minimizing the error in Equation 12. An initial test showed too low (less than 0.015) diffusion rate caused serious instability in the dynamic systems, regardless of the attractor type, setting the lower bound of diffusion rate.

The following Table 4 show the comparison of minimum and maximum error cases between DDM4AP with exponential attractors (DDM-E) and DDM4AP with Gaussian attractors (DDM-G), as the result of optimization: SAR and DR represent "stochastic attraction rate" and "diffusion rate" respectively, used for each case. In

general, DDM-E was better than DDM-G in representing the kurtosis pattern observed in AVA dataset.

Table 4. Comparison of best and worst cases in two dynamic systems models

L2 Error	Exponential	Gaussian
Min	1.38 (SAR=0.1, DR=0.015)	2.08 (SAR=0.2, DR=0.015)
Max	6.61 (SAR=0.25, DR=0.02)	7.33 (SAR=0.2, DR=0.02)

In Figure 16, the effect of stochastic attraction rate (the x-axis factor) to the L2 error is sparsely illustrated for various diffusion rates not less than 0.015: as depicted in the plot, exponential attractors are better than Gaussian counterparts from the view of controllability.

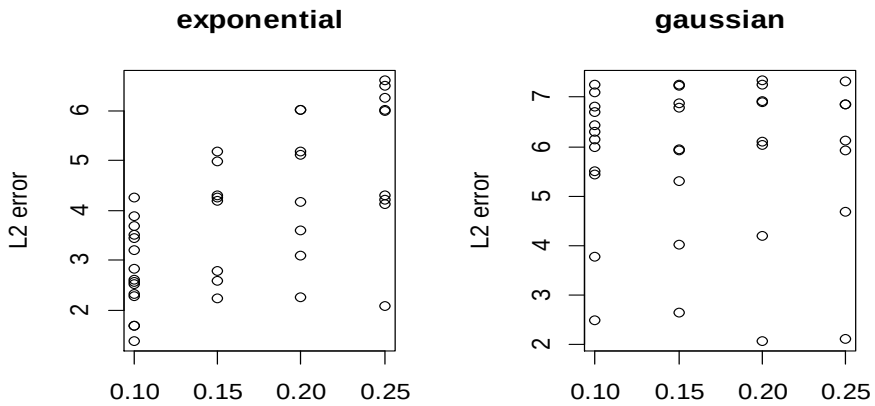


Figure 16. L2 error in accordance with stochastic attraction rates in DDM-E and DDM-G.

The following Table 5 summarizes the four tested models from the view of accordance with the four observed patterns found in AVA dataset and one additional rule (Klaassen boundary). The table reveals the relatively high potential of dynamic system models rather than static models for explaining all the observed patterns.

Table 5. Comparison between models

Requirement	Gaussian	Beta	DDM-G	DDM-E
Convergence to power law in extremes	Pass	Pass	Pass	Pass
Within Klaassen boundary	Pass	Pass	Pass	Pass
Tag-specific effect	Fail	Pass	Pass	Pass
Consensus asymmetry	Fail	Pass	Pass	Pass
Wide kurtosis range	Fail	Fail	Pass	Pass

[a]*DDM-G is DDM4AP with Gaussian pdf while DDM-E with exponential pdf.*

4.4. Discussion

DDM4AP assumes that attractors affect the decision making at various temporal moments. For an instance, an attractor requiring semantic processing (e.g., facial beauty) should involve later than another attractor related with spatial layout (e.g., rule of thirds, golden ratio). Also, because the attractive gain from an attractor is the result of its own neural network – another dynamic system of multiple components, it is hardly expected that the probabilistic distribution of the gain would follow Gaussian

distribution; rather it is likely to follow distributions from power law (e.g., exponential distribution). The more plausible result from the exponential attractors in Figure 15 than that from the Gaussian attractors in Figure 14 is believed to simulate the property as predicted.

Another assumption of DDM4AP regarding visual aesthetic perception as a process of appreciation and decision making (its ancestor, DDM, is originated from the effort of explaining latency distribution in the 2AFC tasks (Ratcliff & McKoon, 2008)) are in accordance with the recent reports from neuroscientific studies. The Orbitofrontal Cortex (OFC) is known to be activated during emotional or aesthetic experience and decision making (Bechara, Damasio, & Damasio, 2000). The left prefrontal dorsolateral cortex (PDC), another region related with decision making, is reported to activate at a latency of 400-1000ms following corresponding activation at 130ms in the visual cortex (Camilo J. Cela-Conde et al., 2004). The recent studies (Camilo J Cela-Conde et al., 2013; Vessel, Starr, & Rubin, 2013) additionally point out the Default Mode Network (DMN) as a shared region between moral and aesthetic appreciation.

Lastly, in the frame of DDM4AP, various interpersonal differences can be simulated by controlling parameters. For example, interpersonal difference of stochastic attraction rates for the same attractors might explain the difference between

amateurs and professional photographers; e.g., the stochastic attraction rate of the professional can be stronger than that of the amateur, owing to training or inborn traits. Inconsistency, the notorious property of aesthetic evaluation, can be simulated by increasing the diffusion rate, W in Equation 11: in DDM4AP, increasing the ratio of a diffusion rate to a mean stochastic attraction rate reinforces the effect of randomness in aesthetic evaluation process, rather than that of given attractor-dependent traits. In the same vein, combination of weak stochastic attraction rates and relatively large diffusion rate might render a character of indecisive and capricious aesthetic appreciation. Personal confidence margin can also be explained by difference in the distance to the rightmost border in Figure 13. Lastly, even though this paper concentrates on early appreciation of visual stimuli and thereby assumes the same number and positions of attractors for each visual stimulus for clarifying the concept of the new model, DDM4AP does not prohibit interpersonal difference caused by difference in aesthetic experience or training which might induce new additional attractors during late appreciation. Interpersonal difference in DDM4AP looks deserving to explore further by analyzing the relation between rating pattern and personal trait (confidence margin, level of aesthetic training, etc.) in the future.

CHAPTER 5. Validation

5.1. Background: Prediction from DDM4AP

As shown in the simulation results in Figure 14 and Figure 15, DDM4AP leads to several predictions about response time (or latency) from the view of a dynamic system as following:

First, the response times of aesthetic evaluation will be significantly different between score groups; specifically, the mediocre and the other (the good or the bad); in other word, evaluating the mediocre should take longer than the good or the bad. This phenomenon was previously reported in another domain, web design appreciation (Tractinsky et al., 2006).

Second, the response time will be significantly affected by the kurtosis of rating distribution; e.g., if diffusion is small enough compared with drift toward an either side, one stimulus with high kurtosis in its rating distribution is expected to have a longer response time than another stimulus with relatively low kurtosis.

To validate the hypotheses induced from DDM4AP about response time, an experiment was conducted to human participants in the following setup.

5.2. Method

For concentrating on response time, tag effect was controlled by collecting stimuli

from single tag group. Considering the relatively less-individual preference on real-world images (Vessel & Rubin, 2010), 100 photos were selected among the 3564 AVA photos of tag1=14("landscape") with more than 100 ratings, and classified into three groups - good, mediocre, and bad - according to the following criteria:

good if median score ≥ 7;

bad if median score ≤ 4; and,

mediocre if median score =5 or 6.

In case of the mediocre group, due to its relatively huge amount (3124 mediocre vs. 216 good vs 204 bad), it is filtered again by the mean and skewness of scores as following:

$5.0 \leq$ mean score ≤ 5.5 ; and,

$-0.1 <$ skewness of scores < 0.1.

Finally, for each group, the topmost and bottommost photos were selected from the view of score kurtosis: the top15 and the bottom 15 for the good and the bad, and the top 20 and the bottom 20 for the mediocre.

In conclusion, 100 photos were selected as stimuli consisting of three groups in accordance with the combination of median scores and kurtosis of score distribution

(therefore the degree of consensus). For analyzing latency-score relation in individual data, the number of photos was determined by the power analysis for ANOVA test comparing 10 groups (the most fine-grained case of counting each score band) to obtain a power of 0.80 when the effect size is large (0.45) and a significance level of 0.05 is employed; considering the case of applying nonparametric tests, the minimum sample size was adjusted from 90 to 95 by applying asymptotic relative efficiency (ARE) of 0.955 for Kruskal-Wallis test, the nonparametric version of ANOVA, and then rounded up to 100.

Ten male and fifteen female students in the age of 20s and 30s with normal or corrected vision participated in this experiment voluntarily with a small compensation of ten dollars for each person. By the activity level of digital photographing, the subjects are categorized into three groups (daily, weekly, monthly) consisting of 6, 18, and 1 members, respectively.

To simulate the online rating environment in DPChallenge.com, a subject sat in front of a 24 inch LCD monitor which displays a photo (stimulus), shown at the center of the screen in original resolution with gray padding.

The subject was requested to evaluate the aesthetic value of the photo on the screen in the same scale with DPChallenge.com by selecting a score button among 10 choices (from 1 to 10). Once the button was pressed, it proceeded to the next photo and

repeated the same task until all one hundred stimuli were scored. Revisiting the previous photos was not allowed.

During the process, the selected aesthetic score and response time were recorded synchronously while the subjects were not aware of the recording of response time, under the control of PsychoPy software (Peirce, 2008).

5.3. Experimental Results

Considering the inter-personal difference in the range of response time, the relation between response time and score was firstly investigated individually for each subject as shown in Figure 17. In Figure 17, a subplot with triangle marks represents a subject whose response times are significantly affected by score (with a 95 percent confidence interval): out of 25 subjects, 11 subjects (44 percent) were classified as significant. However, for each rater, the effect of the average response time was not significant to the score.

For analyzing the general relation between response time and aesthetic score across subjects, quantile normalization was applied to adjust interpersonal difference in response time because of high non-Gaussianity including several outlier which are suspected as the result of temporary attention failure. Then, Kruskal-Wallis rank sum test was applied to see whether or not score affects response time as predicted,

followed by the result of p-value = 4.279e-15 saying that score significantly affect response time with a 95 percent confidence interval. With the same setting, another prediction of significant effect of kurtosis of scores to response time was also confirmed by p-value = 4.335e-14.

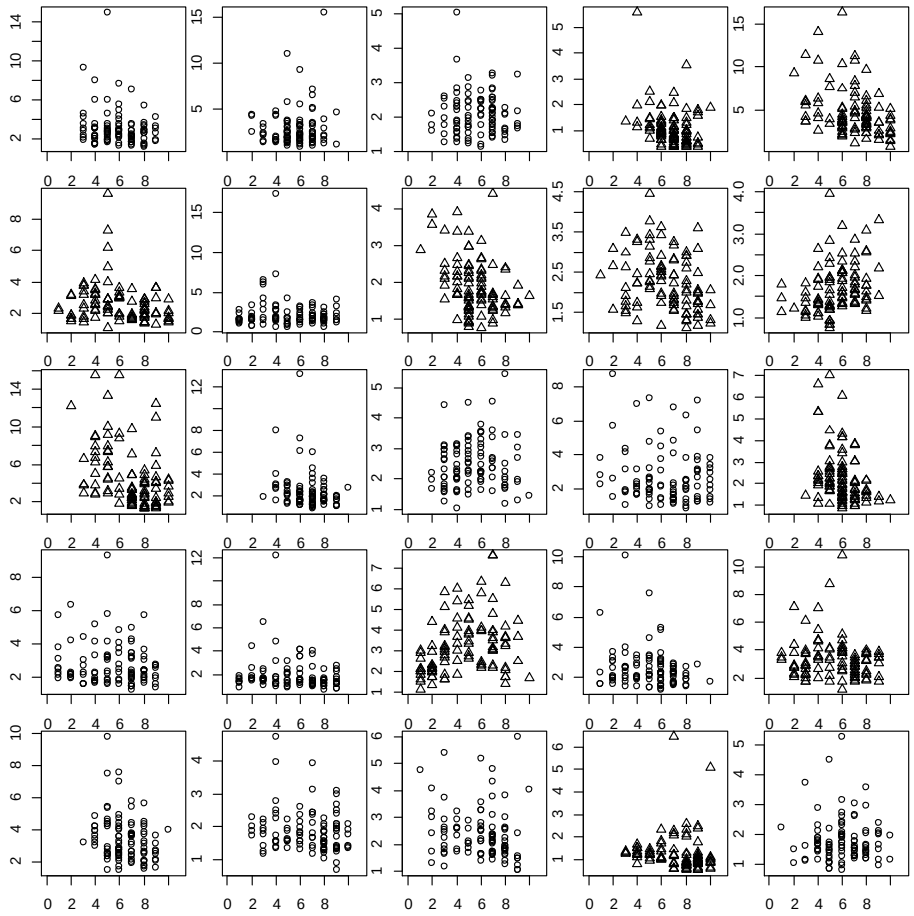


Figure 17. Latency distributions (y-axis in second) per score (x-axis) for 25 subjects. A triangle mark for the subjects whose response time is significantly affected by score, while a circle mark for the other (with a 95 percent confidence interval).

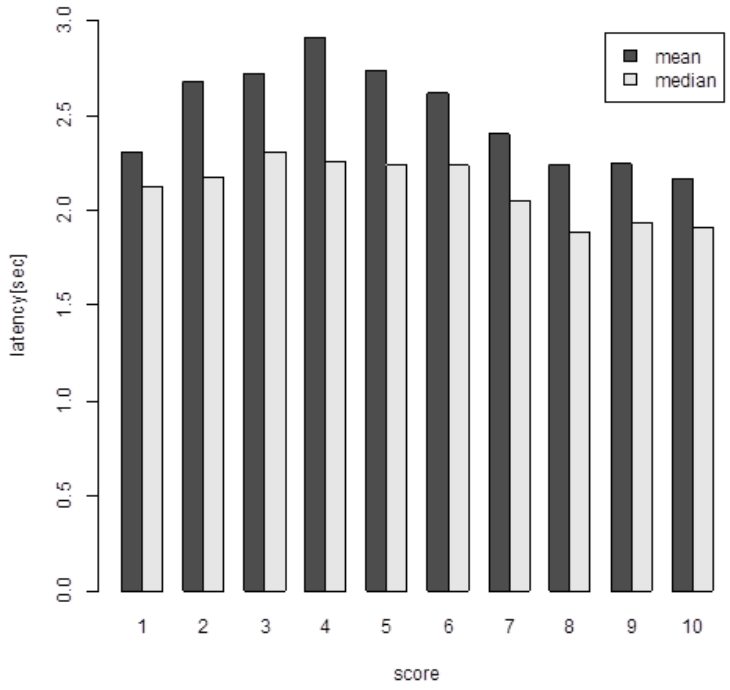


Figure 18. Response times as a function of aesthetic scores

Figure 18 depicts the relation between latency (response time) and aesthetic score. As predicted, the mean response time is longer in the mediocre than in the other groups. The second pattern, asymmetry between the good and the bad, is observed in response time comparison.

Table 6 is the result of pairwise post-hoc comparison using Wilcox-Mann-Whitney rank sum test with the Bonferroni correction. In this test, the confidence interval of 0.95 was adjusted to 0.967, which support the above validation in more quantitative manner. In the pairwise three-class comparison, there is a significant difference

between the mediocre and the good (p-value = 0.00011) while the difference between the mediocre and the bad is not significant (p-value = 0.06826). The power of the pairwise test was 1.0 for all pairs with the significance level of 0.05 and asymptotic relative efficiency of 0.955.

Table 6. Pairwise Post-Hoc Comparisons (Wilcox-Mann-Whitney Rank Sum Test with a "Not Equal" Alternative Hypothesis) Between Response Times of Three Groups

	Bad	**Mediocre**	**Good**
Bad	N/A	0.06826	1e-05
Mediocre	0.06826	N/A	0.00011
Good	1e-05	0.00011	N/A

While latency was significantly affected by the kurtosis of scores for each stimulus (p-value = 4.335e-14 with a 95 percent confidence interval), kurtosis of scores for the mediocre stimuli having 5 as their median sore was not significant for affecting latency (p-value = 0.3624). It is regarded as supporting the second prediction of DDM4AP as a persuasive model for visual aesthetic evaluation because, in the frame of DDM4AP, this result can be explained as the "timeout" for a mediocre stimulus is determined when a particle reaches the rightmost bound, not by the net drift time of the particle during the evaluation process.

One thing to underline is that adjusting the personal difference in latency is a quite difficult task even with quantile normalization because DDM4AP implies the rating

pattern of a subject also affects his/her latency: e.g., uniform vs. bipolar vs. normal (K.-W. Park, Kim, Park, & Zhang, 2014). Therefore, it would be better to emphasize on the intra-personal relation between latency and score within samples from an individual, although the collapsed data are useful for understanding and explaining patterns.

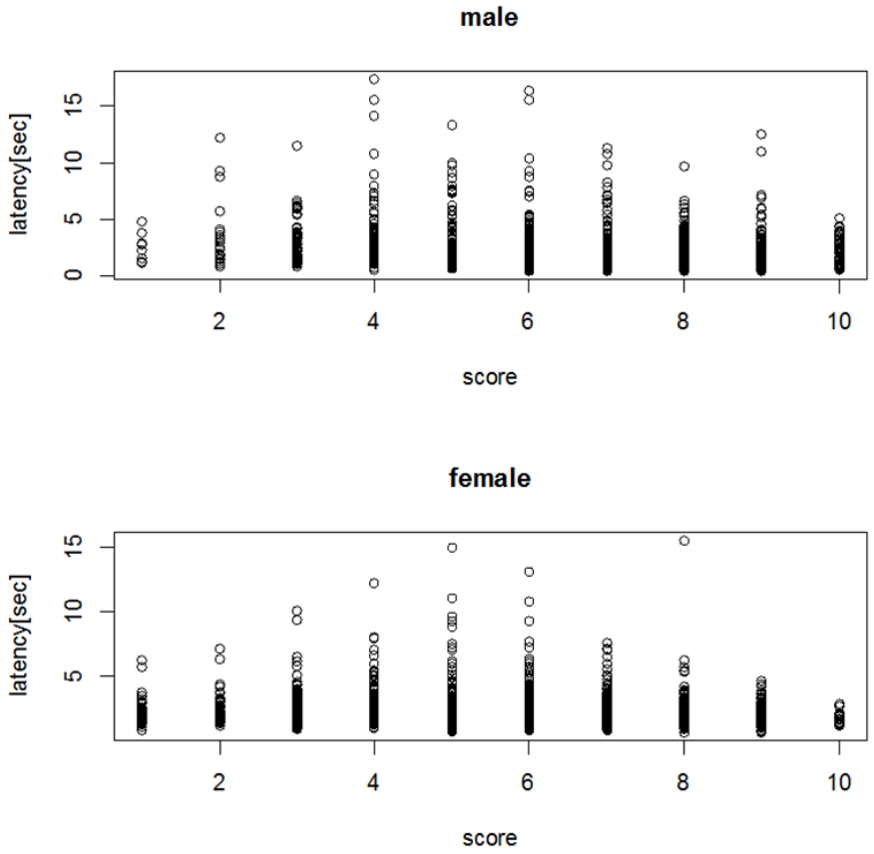


Figure 19. Latency distributions from two gender groups

In additional factor analysis, gender is proven as a significant factor for explaining both response time and score (p-value = 0.000136 and 2.904e-12 respectively). Figure 19 shows the latency distribution of male and female subjects. The effect of mediocrity

to latency is stronger in the female group than in the male group, while the male group shows more dispersed scoring pattern. Although such gender effect can be explained and simulated by difference in the confidence level, numbers and positions of attractors, and stochastic attraction rates in DDM4AP, a more elaborated experiment design with more subjects is required to find out a key factor which is responsible for the gender difference.

Lastly, the correlation between score and the activity level of photographing was analyzed, as the subject group consists of six "daily" photographers, eighteen "weekly" and one "monthly": A question, "How often do you photograph? Daily, weekly, or monthly?" was given while collecting participants via online. The effect of the two activity levels, the daily and the weekly (the monthly is ignored due to too small sample size), was analyzed to score distribution of each photo and counted how many photos show significant difference in sore distribution between the daily and the weekly groups. Kruskal-Wallis rank sum test showed that, with a 95 percent confidence interval, only 10 out of 100 photos were significantly affected by the activity level in their score distribution; with a 99 percent confidence interval, only 2 out of 100 were significant.

For the issue that the reported discovery might be biased to the group of "hobbyists and professionals" (Murray, Marchesotti, & Perronnin, 2012), even though

an additional human participant test with massive number of professionals is required for complete validation as a future work, currently it is estimated from the within-amateur analysis that the found patterns will be consistent with various degrees from amateurs through professionals.

5.4. Discussion

The experimental results support the two hypotheses from DDM4AP that response time is significantly affected by score group and kurtosis, the degree of consensus. The former one is previously reported in the research with the stimuli of web page design (Tractinsky et al., 2006) with a few differences: in the experiments, the response time for the bad photos is longer than the good. it is regarded as owing to the difference in stimuli and questionnaire design, requesting further experiments under more controlled environment for valid comparison including a comparative study.

For the individual analysis result, the type of raters might affect the result as score distributions are different among raters. In other word, if a rater's scoring is deviated from the average rating pattern or distorted by outliers due to the lack of attention, the correlation between response time and score is also distorted. It can be resolved by increasing the number of raters for each stimulus to the level of hundreds as AVA dataset provides.

One important aspect of DDM4AP is that it is able to simulate the translation between discrete emotion models and dimensional models; for an instance, valence can be interpreted as the result of dynamic interaction between several discrete emotion attractors as DDM4AP successfully visualizes.

The difference between mean and median response time implies there are two different mechanisms behind the process. In the frame of DDM4AP, drift-oriented decision and diffusion-oriented decision can explain the duality.

Another issue that DDM4AP raises is that previous "stationary" machine learning models might be limited in their ability of simulating visual aesthetic perception (and, further, emotion), because they don't consider the properties of dynamic systems. If the visual aesthetic perception is the result of a dynamic system including multiple attractors as modeled in DDM4AP, the mixture of positive and negative attractors in a training sample would misguide most machine learning methods unless they have a reliable active learning scheme, another difficult issue in the field of machine learning. It is due to the common assumption of one-to-one or many-to-one relationships, with noisy variance, between samples and their classes behind most machine learning methods, while DDM4AP allows one-to-many relationship additionally and regards "variance", or even "bifurcation", as a systematic consequence. The pioneering work of Wu et al. (O. Wu et al., 2011) using SVRD has potential of making synergy with

DDM4AP. For an instance, DDM4AP might help SVRD treat samples with null stimuli, which include none of attractor, differently in the mediocre group, or adjust weights of score bands during training based on the level of consensus. Although SVRD didn't show a comparable performance in 9,000 photos from dpchallenge.com (the fraction of AVA dataset named as "DS2" in (O. Wu et al., 2011)) to the good result with Photo.Net ("DS1" in (O. Wu et al., 2011)), it is believed that their approach of multi-label classification is fundamentally correct and promising, hoping to expand Wu's work in the future work of application: that will be a combination of a dynamic model in psychology and multi-label (1-to-N) classification in machine learning.

This explains why the pragmatic solution of excluding low consensus samples during the training stage (Ke et al., 2006; Murray, Marchesotti, & Perronnin, 2012) was so effective to enhance the performance of traditional classifiers such as support vector machines or Bayesian classifiers in the task of aesthetics evaluation from low-level features. In the same vein, the usual practice of excluding images in the mediocre group during training should be also effective because the stimuli are the most likely to have multiple and balanced (between the positive and the negative) attractors, or nothing as a null stimulus; either case is sufficient to prevent classifiers from being learned.

From the view of computer vision, the predicted presence of two types of the

mediocrity in DDM4AP implies that they should be treated differently to each other when finding features or training classifiers. For example, if a set of one-vs-one classifiers are used for multi-label classification, one mediocre sample without any like or dislike factor should be treated differently to the other mediocre sample having equal numbers of like and dislike factors. In other case, it would be more promising to construct a multi-label classifier from a set of one-vs-all classifiers if it is possible to detect and exclude the mediocre sample of "balanced-between-two-attractor-groups" before training, because it is unlikely that the two types share a similar embedding pattern in feature space.

Lastly, the correlation analysis between the rating pattern and activity level of photographing alleviates the concern that there might be an intrinsic sample bias because the data were collected from people willing to post their photos and rate other's works. According to the AVA providers, the ground truth data in community-based online photo evaluation are from multiple raters who are generally "prosumers of data"(Murray, Marchesotti, Perronnin, et al., 2012). AVA providers insist that scale mitigates it as following:

> *…..Each image is associated with a distribution of scores which correspond to individual votes. The number of votes per image ranges from 78 to 549, with an average of 210 votes. Such score distributions represent a gold mine of aesthetic judgments generated by hundreds of*

amateur and professional photographers with a practiced eye. We believe that such annotations have a high intrinsic value because they capture the way hobbyists and professionals understand visual aesthetics.

Because AVA creators don't know the exact ratio of hobbyists to professionals due to the lack of identity check in online voting systems, the very mechanism which enables massive participation, it is logically correct to say that the found pattern is validated only in the group of hobbyists and professionals.

Conclusion

Studies of visual aesthetic perception from various disciplines across philosophy, psychology, neuroscience, and computer science were reviewed for better understanding of the nature of aesthetic judgement in human mind and thereby providing valuable intuitions to computational aesthetics.

A machine-learning-based aesthetic value estimator was implemented with a new descriptor named LoSC for capturing spatial information, resulting in comparable performance with the state of art in this field.

Motivated by the abnormal effect of the mediocre photos to performance in computational aesthetics, consensus analysis was performed to the massive aesthetics dataset, reporting the finding of four patterns behind the score distributions: wide kurtosis range, consensus asymmetry, the 4/3 power law regime at both extremes, and tag effects.

Because a simple probabilistic distribution model (e.g., a unimodal Gaussian distribution) is inadequate to explain or simulate these patterns, several alternative models of visual aesthetic perception were proposed and evaluated by the representation of the observed patterns in their simulation results, concluding that a dynamic model named DDM4AP, a modification of drift-diffusion model, is most successful for simulating all the patterns owing to its mechanism of determining the

perceived aesthetic value by the spatiotemporal interaction between multiple attractors with random noise.

To evaluate the feasibility of DDM4AP as a model of visual aesthetic perception in human mind, its innate property of dependency between perceived aesthetic values and their response times was tested via a human participant experiment. The experimental results show that the dependency exists as DDM4AP predicted, supporting the model as reflecting core properties of visual aesthetic evaluation process.

One of the benefits that DDM4AP provides is that it is appropriate to explain the mixed nature of aesthetic judgement, a mixture of spontaneous perception and automated evaluation in the frame of decision making. Considering the merit of DDM4AP in simulating decision making as the result of interactions among contrastive attractors, it seems deserving to expand the application of the DDM4AP to more general decision making or emotion involving a reward system beyond aesthetic appreciation discussed so far. For example, optimal selection among options having their own benefits and costs as a pair is appropriate to be modeled by DDM4AP. In the more related domain of emotion, DDM4AP might be able to explain "mixed emotion" as a natural consequence from balanced contrastive emotional percepts.

It would be desirable if the proposed approach helps giving intuition for breaking the "glass-ceiling" (Pachet & Aucouturier, 2004), which has been regarded as the

consequence of the semantic gap between features and high-level perception, in prediction of emotion. For an instance, it might deserves to try an active learning model for evaluating visual aesthetics of photos by adjusting learning rate or changing combination in ensemble methods according to the result of sophisticated consensus analysis on the rating pattern. Also, this paper raises an issue of developing a new machine learning concept which is able to treat dynamic processes properly. For example, in the context of the current multi-class classification, if one of the classes has various consensus level and therefore being suspected as the interaction between two contrastive factors, the training of the classifiers is better to adopt "1-vs-all" strategy rather than the "1-vs-1" for minimizing the effect of the misleading samples in the less-agreed class. In the same vein, regression will be more robust to the presence of the less-agreed samples than classification.

References

Aharon, I., Etcoff, N., Ariely, D., Chabris, C. F., O'Connor, E., & Breiter, H. C. (2001). Beautiful faces have variable reward value: fMRI and behavioral evidence. *Neuron, 32*(3), 537-551.

Arnheim, R. (1954). *Art and visual perception*: Univ of California Press.

Atalay, B. (2004). Math and the Mona Lisa. *The Art and Science of Leonardo da Vinci*.

Balanda, K. P., & MacGillivray, H. (1988). Kurtosis: a critical review. *The American Statistician, 42*(2), 111-119.

Balling, J. D., & Falk, J. H. (1982). Development of visual preference for natural environments. *Environment and Behavior, 14*(1), 5-28.

Bar, M., & Neta, M. (2006). Humans prefer curved visual objects. *Psychological Science, 17*(8), 645-648.

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex, 10*(3), 295-307.

Becker-Asano, C., & Wachsmuth, I. (2008). Affect simulation with primary and secondary emotions. In *8th International Conference on Intelligent Virtual Agents, 2008, IVA2008*.

Benson, P. L., Karabenick, S. A., & Lerner, R. M. (1976). Pretty pleases: The effects of physical attractiveness, race, and sex on receiving help. *Journal of Experimental Social Psychology, 12*(5), 409-415.

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York: Appleton, 1971.

Berridge, K. C., & Kringelbach, M. L. (2008). Affective neuroscience of pleasure: reward in humans and animals. *Psychopharmacology, 199*(3), 457-480.

Bertamini, M., Bennett, K. M., & Bode, C. (2011). The anterior bias in visual art: The case of images of animals. *Laterality: Asymmetries of Body, Brain and Cognition, 16*(6), 673-

689.

Bhattacharya, S., Nojavanasghari, B., Chen, T., Liu, D., Chang, S.-F., & Shah, M. (2013). *Towards a comprehensive computational model foraesthetic assessment of videos*.

Bhattacharya, S., Sukthankar, R., & Shah, M. (2011). A holistic approach to aesthetic enhancement of photographs. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 7*(1), 21.

Biederman, I., & Vessel, E. (2006). Perceptual pleasure and the brain, *American Scientist, 94*(3), 247-253.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700.

Borth, D., Chen, T., Ji, R., & Chang, S.-F. (2013). SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM International Conference on Multimedia*, 459-460.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Burke, E. (1796). *On the sublime and beautiful*: BiblioBytes.

Burke, E. (1812). *A philosophical enquiry into the Origin of our ideas of the sublime and beautiful*: FC and J. Rivington, Otridge and son.

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain sciences, 12*(01), 1-14.

Butcher, S. H. (1951). *Aristotle's theory of poetry and fine art: with a critical text and translation of the Poetics. With a prefatory essay, Aristotelian literary criticism* (Vol. 42): Courier Corporation.

Cavanagh, P. (2005). The artist as neuroscientist. *Nature, 434*(7031), 301-307.

Cela-Conde, C. J., García-Prieto, J., Ramasco, J. J., Mirasso, C. R., Bajo, R., Munar, E., and Maestú, F. (2013). Dynamics of brain networks in the aesthetic appreciation. *Proceedings of the National Academy of Sciences, 110*(Supplement 2), 10454-10461.

Cela-Conde, C. J., Marty, G., Maestú, F., Ortiz, T., Munar, E., Fernández, A., and Quesney, F. (2004). Activation of the prefrontal cortex in the human visual aesthetic perception. *Proceedings of the National Academy of Sciences of the United States of America, 101*(16), 6321-6325.

Cerosaletti, C. D., & Loui, A. C. (2009). Measuring the perceived aesthetic quality of photographic images. In *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on* (47-52). IEEE,

Chatterjee, A. (2004). Prospects for a cognitive neuroscience of visual aesthetics. *Bulletin of Psychology and the Arts, 4*(2), 56-60.

Chatterjee, A. (2006). The neuropsychology of visual art: Conferring capacity. *International Review of Neurobiology, 74*, 39.

Chatterjee, A. (2011). Neuroaesthetics: a coming of age story. *Journal of Cognitive Neuroscience, 23*(1), 53-62.

Chatterjee, A., Hamilton, R. H., & Amorapanth, P. X. (2006). Art produced by a patient with Parkinson's disease. *Behavioural Neurology, 17*(2), 105-108.

Chatterjee, A., Thomas, A., Smith, S. E., & Aguirre, G. K. (2009). The neural response to facial attractiveness. *Neuropsychology, 23*(2), 135.

Chatterjee, A., Widick, P., Sternschein, R., Smith, W. B., & Bromberger, B. (2010). The assessment of art attributes. *Empirical Studies of the Arts, 28*(2), 207-222.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*,. 321-

357.

Ciesielski, V., Barile, P., & Trist, K. (2013). *Finding image features associated with high aesthetic value by machine learning*: Springer Berlin Heidelberg.

Conway, B. R., & Livingstone, M. S. (2007). Perspectives on science and art. *Current Opinion in Neurobiology, 17*(4), 476-482.

Cristelli, M., Zaccaria, A., & Pietronero, L. (2012). Universal relation between skewness and kurtosis in complex dynamics. *Physical Review E, 85*(6), 066108.

Cunningham, M. R., Barbee, A. P., & Philhower, C. L. (2002). Dimensions of facial physical attractiveness: The intersection of biology and culture.

Cunningham, M. R., Roberts, A. R., Barbee, A. P., Druen, P. B., & Wu, C.-H. (1995). " Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology, 68*(2), 261.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on,* 886-893. IEEE.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006,* 288-301. Springer Berlin Heidelberg.

Datta, R., & Wang, J. Z. (2010). ACQUINE: aesthetic quality inference engine-real-time automatic rating of photo aesthetics. In *Proceedings of the International Conference on Multimedia Information Retrieval,* 421-424. ACM.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*(3), 285.

Ditterich, J. (2010). A comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. *Frontiers in Neuroscience, 4*.

Dutton, D. (2003). Aesthetics and evolutionary psychology. *The Oxford Handbook for Aesthetics*, 693-705.

Eysenck, H. J. (1940). THE GENERAL FACTOR IN AESTHETIC JUDGEMENTS1. *British Journal of Psychology. General Section, 31*(1), 94-102.

Fechner, G. T. (1876). *Vorschule der aesthetik* (Vol. 1): Breitkopf & Härtel.

Fernandez, D., & Wilkins, A. J. (2008). Uncomfortable images in art and nature. *Perception, 37*(7), 1098.

Freud, S., Strachey, J., Cixous, H., & Dennomé, R. (1976). Fiction and its phantoms: a reading of Freud's das unheimliche (the" uncanny"). *New Literary History*, 525-645.

Galanter, P. (2012). Computational aesthetic evaluation: past and future. *Computers and Creativity*, 255-293. Springer.

Gangestad, S. W., & Buss, D. M. (1993). Pathogen prevalence and human mate preferences. *Ethology and Sociobiology, 14*(2), 89-96.

Gibson, J. J. (1950). *The perception of the visual world.* Oxford, England: Houghton Mifflin, xii, 242.

Gill, M. B. (2011). Lord Shaftesbury [Anthony Ashley Cooper, 3rd Earl of Shaftesbury]. *The Stanford Encyclopedia of Philosophy.* Fall 2011.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London B: Biological Sciences, 205*(1161), 581-598.

Graham, D. J., & Field, D. J. (2008). Statistical regularities of art images and natural scenes:

Spectra, sparseness and nonlinearities. *Spatial Vision, 21*(1-2), 149-164.

Grammer, K., Fink, B., Møller, A. P., & Thornhill, R. (2003). Darwinian aesthetics: sexual selection and the biology of beauty. *Biological Reviews, 78*(3), 385-407.

Grammer, K., & Thornhill, R. (1994). Human (Homo sapiens) facial attractiveness and sexual selection: the role of symmetry and averageness. *Journal of Comparative Psychology, 108*(3), 233.

Griffith, T., & Ferrari, G. R. F. (2000). *Plato:'The Republic'*: Cambridge University Press.

Guilford, J. P., & Smith, P. C. (1959). A system of color-preferences. *The American Journal of Psychology*, 487-502.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Han, K.-T. (2010). An exploration of relationships among the responses to natural scenes scenic beauty, preference, and restoration. *Environment and Behavior, 42*(2), 243-270.

Hanjalic, A., & Xu, L.-Q. (2001). User-oriented affective video content analysis. In *Content-Based Access of Image and Video Libraries, 2001.(CBAIVL 2001). IEEE Workshop on*, 50-57. IEEE.

He, K., Sun, J., & Tang, X. (2011). Single image haze removal using dark channel prior. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33*(12), 2341-2353.

Hekkert, P., & Van Wieringen, P. C. W. (1996). Beauty in the eye of expert and nonexpert beholders: A study in the appraisal of art. *The American Journal of Psychology*, 389-407.

Horvath, T. (1979). Correlates of physical beauty in men and women. *Social Behavior and Personality: an International Journal, 7*(2), 145-151.

Hume, D. (2000). *(Of The) Standard of Taste*: BiblioBytes.

Hurlbert, A. C., & Ling, Y. (2007). Biological components of sex differences in color preference. *Current Biology, 17*(16), R623-R625.

Hutcheson, F. (1729). *An Inquiry Into the Original of Our Ideas of Beauty and Virtue: In Two Treatises, The 2nd Ed., corrected and enlarg'd*.

Ishai, A., Fairhall, S. L., & Pepperell, R. (2007). Perception, memory and aesthetics of indeterminate art. *Brain Research Bulletin, 73*(4), 319-324.

Jackson, L. A., & Ervin, K. S. (1992). Height stereotypes of women and men: The liabilities of shortness for both sexes. *The Journal of Social Psychology, 132*(4), 433-445.

Jacobsen, T. (2006). Bridging the arts and sciences: A framework for the psychology of aesthetics. *Leonardo, 39*(2), 155-162.

Jacobsen, T., & Hofel, L. (2002). Aesthetic judgments of novel graphic patterns: analyses of individual judgments. *Perceptual and Motor skills, 95*(3), 755-766.

Jahoda, G. (2005). Theodor Lipps and the shift from "sympathy" to "empathy". *Journal of the History of the Behavioral Sciences, 41*(2), 151-163.

Jones, D., & Hill, K. (1993). Criteria of facial attractiveness in five populations. *Human Nature, 4*(3), 271-296.

Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q.-T., Wang, J. Z., Li, J., & Luo, J. (2011). Aesthetics and emotions in images. *Signal Processing Magazine, IEEE, 28*(5), 94-115.

Juneja, M., Vedaldi, A., Jawahar, C., & Zisserman, A. (2013). Blocks that shout: distinctive parts for scene classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 923-930. IEEE.

Kampe, K. K. W., Frith, C. D., Dolan, R. J., & Frith, U. (2001). Psychology: Reward value of attractiveness and gaze. *Nature, 413*(6856), 589-589.

Kant, I. (1952). *The critique of judgement, trans. JC Meredith*. Oxford: Oxford University Press, 314, 175-176.

Kaplan, R., Kaplan, S., & Brown, T. (1989). Environmental preference a comparison of four domains of predictors. *Environment and Behavior, 21*(5), 509-530.

Kaplan, S., Kaplan, R., & Wendt, J. S. (1972). Rated preference and complexity for natural and urban visual material. *Perception & Psychophysics, 12*(4), 354-356.

Kawabata, H., & Zeki, S. (2004). Neural correlates of beauty. *Journal of Neurophysiology, 91*(4), 1699-1705.

Ke, Y., Tang, X., & Jing, F. (2006). The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on,* 1, 419-426. IEEE.

Kelso, J. S. (1997). *Dynamic patterns: The self-organization of brain and behavior*: MIT press.

Kenealy, P., Frude, N., & Shaw, W. (1988). Influence of children's physical attractiveness on teacher expectations. *The Journal of Social Psychology, 128*(3), 373-383.

Klaassen, C. A., Mokveld, P. J., & Van Es, B. (2000). Squared skewness minus kurtosis bounded by 186/125 for unimodal distributions. *Statistics & Probability Letters, 50*(2), 131-135.

Kobayashi, S. (1981). The aim and method of the color image scale. *Color Research & Application, 6*(2), 93-107.

Konečni, V. J., & Cline, L. E. (2001). The" golden woman:" an exploratory study of women's proportions in paintings. *Visual Arts Research*, 69-78.

Konkle, T., & Oliva, A. (2011). Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 23.

Kranz, F., & Ishai, A. (2006). Face perception is modulated by sexual preference. *Current Biology, 16*(1), 63-68.

Kwon, J. S., Kim, J.-J., Lee, D. W., Lee, J. S., Lee, D. S., Kim, M.-S., . . . Lee, M. C. (2003). Neural correlates of clinical symptoms and cognitive dysfunctions in obsessive–compulsive disorder. *Psychiatry Research: Neuroimaging, 122*(1), 37-47.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*(3), 390.

Langlois, J. H., Ritter, J. M., Roggman, L. A., & Vaughn, L. S. (1991). Facial diversity and infant preferences for attractive faces. *Developmental Psychology, 27*(1), 79.

Langlois, J. H., Roggman, L. A., & Rieser-Danner, L. A. (1990). Infants' differential social responses to attractive and unattractive faces. *Developmental Psychology, 26*(1), 153.

Latto, R., Brain, D., & Kelly, B. (2000). An oblique effect in aesthetics: Homage to Mondrian (1872-1944). *Perception, 29*(8), 981-988.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR), 2006 IEEE Computer Society Conference on, 2,* 2169-2178. IEEE.

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95*(4), 489-508.

Leder, H., Tinio, P. P., & Bar, M. (2011). Emotional valence modulates the preference for curved objects. *Perception, 40*(6), 649.

Leder, H., Tinio, P. P. L., Fuchs, I. M., & Bohrn, I. (2010). When attractiveness demands longer looks: The effects of situation and gender. *The Quarterly Journal of Experimental Psychology, 63*(9), 1858-1871.

Lerner, R. M., Lerner, J. V., Hess, L. E., Schwab, J., Jovanovic, J., Talwar, R., & Kucher, J. S.

(1991). Physical attractiveness and psychosocial functioning among early adolescents. *The Journal of Early Adolescence, 11*(3), 300-320.

Leyssen, M. H., Linsen, S., Sammartino, J., & Palmer, S. E. (2012). Aesthetic preference for spatial composition in multiobject pictures. *i-Perception, 3*(1), 25.

Lim, S.-L., O'Doherty, J. P., & Rangel, A. (2013). Stimulus value signals in ventromedial PFC reflect the integration of attribute value signals computed in fusiform gyrus and posterior superior temporal gyrus. *The Journal of Neuroscience, 33*(20), 8729-8741.

Linsen, S., Leyssen, M. H., Gardner, J. S., & Palmer, S. E. (2010). Aesthetic preferences in the size of images of real-world objects. *Journal of Vision, 10*(7), 1234-1234.

Lipps, T. (1935). Empathy, inner imitation, and sense-feelings. *A modern book of aesthetics,* New York: Holt and Company.(Original work published 1903).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110.

Mai, L., Le, H., Niu, Y., & Liu, F. (2011). Rule of thirds detection from photograph. In *Multimedia (ISM), 2011 IEEE International Symposium on,* 91-96. IEEE.

Marchesotti, L., Perronnin, F., Larlus, D., & Csurka, G. (2011). Assessing the aesthetic quality of photographs using generic image descriptors. In *Computer Vision (ICCV), 2011 IEEE International Conference on,* 1784-1791. IEEE.

Marr, D. (1982). *Vision : a computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

McManusU, I., Jones, A. L., & Cottrell, J. (1981). The aesthetics of colour. *Perception, 10*(6), 651-666.

Mealey, L., Bridgstock, R., & Townsend, G. C. (1999). Symmetry and perceived facial attractiveness: a monozygotic co-twin comparison. *Journal of Personality and Social*

Psychology, 76(1), 151.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*(1), 89-115.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27*(10), 1615-1630.

Miller, B. L., & Hou, C. E. (2004). Portraits of artists: emergence of visual creativity in dementia. *Archives of Neurology, 61*(6), 842-844.

Miller, G. F. (1999). Sexual selection for cultural displays. *The Evolution of Culture*, 71-91.

Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2408-2415. IEEE.

Murray, N., Marchesotti, L., Perronnin, F., & Meylan, F. (2012). Learning to rank images using semantic and aesthetic labels. In *British Machine Vision Conference (BMVC)*, 1-10.

Møller, A. P. (1992). Female swallow preference for symmetrical male sexual ornaments. *Nature, 357*(6375), 238-240.

Møller, A. P., & Swaddle, J. P. (1997). *Asymmetry, developmental stability and evolution*: Oxford University Press.

Müller, H., Clough, P., Deselaers, T., Caputo, B., & CLEF, I. (2010). Experimental evaluation in visual information retrieval. *The Information Retrieval Series, 32*.

Nadal, M., Munar, E., Capó, M. À., Rossello, J., & Cela-Conde, C. J. (2008). Towards a framework for the study of the neural correlates of aesthetic preference. *Spatial Vision, 21*(3), 379-396.

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature*

Neuroscience, 4(1), 95-102.

Obrador, P., Schmidt-Hackenberg, L., & Oliver, N. (2010). The role of image composition in image aesthetics. In *Image Processing (ICIP), 2010 17th IEEE International Conference on,* 3185-3188. IEEE.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24*(7), 971-987.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34,* 72-107.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155*, 23-36.

Orians, G. H., & Heerwagen, J. H. (1992). Evolved responses to landscapes. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* New York, NY, US: Oxford University Press, xii, 666

Ou, L. C., Luo, M. R., Woodcock, A., & Wright, A. (2004). A study of colour emotion and colour preference. Part III: Colour preference modeling. *Color Research & Application, 29*(5), 381-389.

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia, 41*(2), 147-155.

Pachet, F., & Aucouturier, J.-J. (2004). Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences, 1*(1), 1-13.

Palmer, S. E., & Griscom, W. S. (2013). Accounting for taste: Individual differences in

preference for harmony. *Psychonomic Bulletin & Review, 20*(3), 453-461.

Palmer, S. E., & Schloss, K. B. (2010). An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences, 107*(19), 8877-8882.

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual Review of Psychology, 64*, 77-107.

Park, K.-W., Kim, B.-H., Park, T.-S., & Zhang, B.-T. (2014). Uncovering response biases in recommendation. In *Workshops at the 28th AAAI Conference on Artificial Intelligence*.

Park, T., & Zhang, B. (2015). Consensus analysis and modeling of visual aesthetic perception. *Affective Computing, IEEE Transactions on, PP*(99), 1-1. doi:10.1109/TAFFC.2015.2400151

Park, T.-S., Kim, B.-H., & Zhang, B.-T. (2014). A viewer preference model based on physiological feedback. *Journal of Korean Institute of Intelligent Systems, 24*(3), 7.

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*.

Pelowski, M., & Akiba, F. (2011). A model of art perception, evaluation and emotion in transformative aesthetic experience. *New Ideas in Psychology, 29*(2), 80-97.

Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature, 368*(6468), 239-242.

Peters, G. (2007). Aesthetic primitives of images for visualization. In *Information Visualization, 2007. IV'07. 11th International Conference*, 316-325. IEEE.

Pettijohn Ii, T. F., & Tesser, A. (1999). Popularity in environmental context: Facial feature assessment of American movie actresses. *Media Psychology, 1*(3), 229-247.

Pettijohn, T. F., & Jungeberg, B. J. (2004). Playboy playmate curves: Changes in facial and body feature preferences across social and economic conditions. *Personality and Social*

Psychology Bulletin, 30(9), 1186-1197.

Picard, R. W. (1997). *Affective computing*: Cambridge: MIT press.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization, *Advances in Kernel Methods-Support Vector Learning, 3*.

Provost, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets*, 1-3.

Provost, M. P., Quinsey, V. L., & Troje, N. F. (2008). Differences in gait across the menstrual cycle and their attractiveness to men. *Archives of Sexual Behavior, 37*(4), 598-604.

Ramachandran, V. S., & Hirstein, W. (1999). The science of art: A neurological theory of aesthetic experience. *Journal of Consciousness Studies, 6*(6-7), 15-51.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873-922.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8*(4), 364-382.

Ricci, F., Rokach, L., & Shapira, B. (2011). *Introduction to recommender systems handbook*: Springer US.

Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions, and judgments of physically attractive students: A review. *Review of Educational Research, 62*(4), 413-426.

Rubenstein, A. J., Kalakanis, L., & Langlois, J. H. (1999). Infant preferences for attractive faces: a cognitive explanation. *Developmental Psychology, 35*(3), 848.

Russell, P. A., & George, D. A. (1990). Relationships between aesthetic response scales applied

to paintings. *Empirical Studies of the Arts, 8*(1), 15-30.

Sachs, T. S., Kakarala, R., Castleman, S. L., & Rajan, D. (2011). A data-driven approach to understanding skill in photographic composition. In *Computer Vision–ACCV 2010 Workshops*, 112-121. Springer Berlin Heidelberg.

Sacks, O. (1995). *An anthropologist on Mars: Seven paradoxical tales by Oliver Sacks*: New York: Knopf.

Sammartino, J., & Palmer, S. E. (2012). Aesthetic issues in spatial composition: Effects of vertical position and perspective on framing single objects. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 865.

Sattin, F., Agostini, M., Cavazzana, R., Serianni, G., Scarin, P., & Vianello, N. (2009). About the parabolic relation existing between the skewness and the kurtosis in time series of experimental data. *Physica Scripta, 79*(4), 045006.

Savakis, A. E., Etz, S. P., & Loui, A. C. (2000). Evaluation of image appeal in consumer photography. In *Electronic Imaging*, 111-120. International Society for Optics and Photonics.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*(4), 695-729.

Senior, C. (2003). Beauty in the brain of the beholder. *Neuron, 38*(4), 525-528.

Shimamura, A. P., & Palmer, S. E. (2012). *Aesthetic science: Connecting minds, brains, and experience*: Oxford University Press.

Silvia, P. J., & Barona, C. M. (2009). Do people prefer curved objects? Angularity, expertise, and aesthetic preference. *Empirical Studies of the Arts, 27*(1), 25-42.

Singer, F., Riechert, S. E., Xu, H., Morris, A. W., Becker, E., Hale, J. A., & Noureddine, M. A. (2000). Analysis of courtship success in the funnel-web spider Agelenopsis aperta.

Behaviour, 137(1), 93-117.

Singh, D. (1993). Adaptive significance of female physical attractiveness: role of waist-to-hip ratio. *Journal of Personality and Social Psychology, 65*(2), 293.

Singh, S., Gupta, A., & Efros, A. A. (2012). Unsupervised discovery of mid-level discriminative patches. In *Computer Vision–ECCV 2012*, 73-86. Springer.

Slater, A., Von der Schulenburg, C., Brown, E., Badenoch, M., Butterworth, G., Parsons, S., & Samuels, C. (1998). Newborn infants prefer attractive faces. *Infant Behavior and Development, 21*(2), 345-354.

Soleymani, M., & Larson, M. (2010). Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, 4-8. ACM.

Solli, M., & Lenz, R. (2010). Color semantics for image indexing. In *Conference on Colour in Graphics, Imaging, and Vision, 1*, 353-358. Society for Imaging Science and Technology.

Sroufe, R., Chaikin, A., Cook, R., & Freeman, V. (1976). The effects of physical attractiveness on honesty: A socially desirable response. *Personality and Social Psychology Bulletin, 3*(1), 59-62.

Su, H.-H., Chen, T.-W., Kao, C.-C., Hsu, W. H., & Chien, S.-Y. (2012). Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. *Multimedia, IEEE Transactions on, 14*(3), 833-843.

Thornhill, R., & Gangestad, S. W. (1994). Human fluctuating asymmetry and sexual behavior. *Psychological Science, 5*(5), 297-302.

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Sciences,*

3(12), 452-460.

Tooby, J., & Cosmides, L. (2001). Does beauty build adapted minds? Toward an evolutionary theory of aesthetics, fiction, and the arts. *SubStance, 30*(1), 6-27.

Tractinsky, N., Cokhavi, A., Kirschenbaum, M., & Sharfi, T. (2006). Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human-Computer Studies, 64*(11), 1071-1083.

Ungerleider, L. G. (1982). Two cortical visual systems. *Analysis of Visual Behavior*, 549-586.

Ursu, S., Stenger, V. A., Shear, M. K., Jones, M. R., & Carter, C. S. (2003). Overactive action monitoring in obsessive-compulsive disorder evidence from functional magnetic resonance imaging. *Psychological Science, 14*(4), 347-353.

Vartanian, O., & Goel, V. (2004). Neuroanatomical correlates of aesthetic preference for paintings. *Neuroreport, 15*(5), 893-897.

Vartanian, O., Navarrete, G., Chatterjee, A., Fich, L. B., Leder, H., Modroño, C., . . . Skov, M. (2013). Impact of contour on aesthetic judgments and approach-avoidance decisions in architecture. *Proceedings of the National Academy of Sciences, 110*(Supplement 2), 10446-10453.

Vessel, E. A., & Rubin, N. (2010). Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of Vision, 10*(2).

Vessel, E. A., Starr, G. G., & Rubin, N. (2013). Art reaches within: aesthetic experience, the self and the default mode network. *Frontiers in Neuroscience, 7*.

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science, 17*(9), 799-806.

Winston, J. S., O'Doherty, J., Kilner, J. M., Perrett, D. I., & Dolan, R. J. (2007). Brain systems for assessing facial attractiveness. *Neuropsychologia, 45*(1), 195-206.

Woods, W. A. (1991). Parameters of aesthetic objects: Applied aesthetics. *Empirical Studies of the Arts, 9*(2), 105-114.

Wu, J., & Rehg, J. M. (2011). CENTRIST: A visual descriptor for scene categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33*(8), 1489-1501.

Wu, O., Hu, W., & Gao, J. (2011). Learning to predict the perceived visual quality of photos. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 225-232. IEEE.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE conference on*, 3485-3492. IEEE.

Yue, X., Vessel, E. A., & Biederman, I. (2007). The neural basis of scene preferences. *Neuroreport, 18*(6), 525-529.

Zabih, R., & Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *Computer Vision—ECCV'94*, 151-158. Springer.

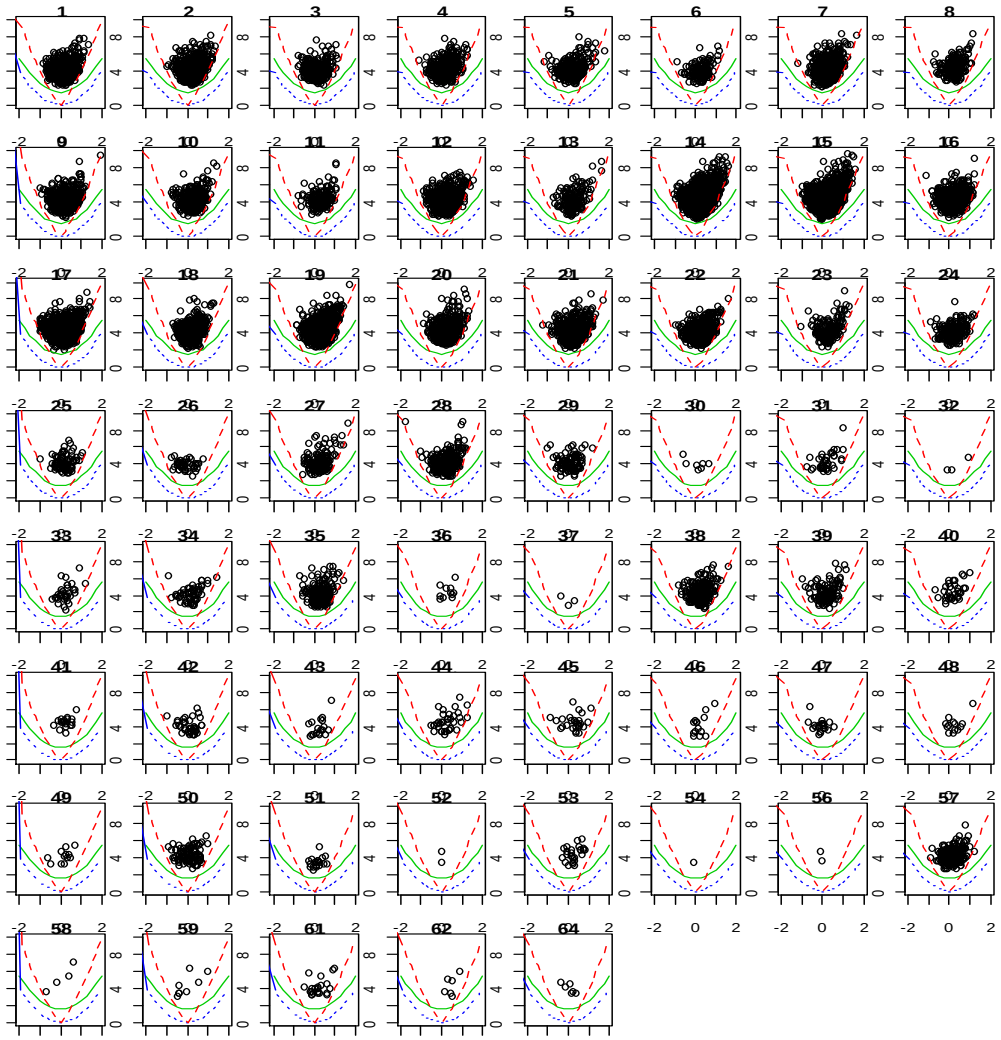
Zaidel, D. (2005). Neuropsychology of art. *Neurological, Cognitive and Evolutionary*.

Zeki, S. (1999). Art and the brain. *Journal of Consciousness Studies, 6*(6-7), 76-96.

Appendix 1. Free vs. Non-Free Study

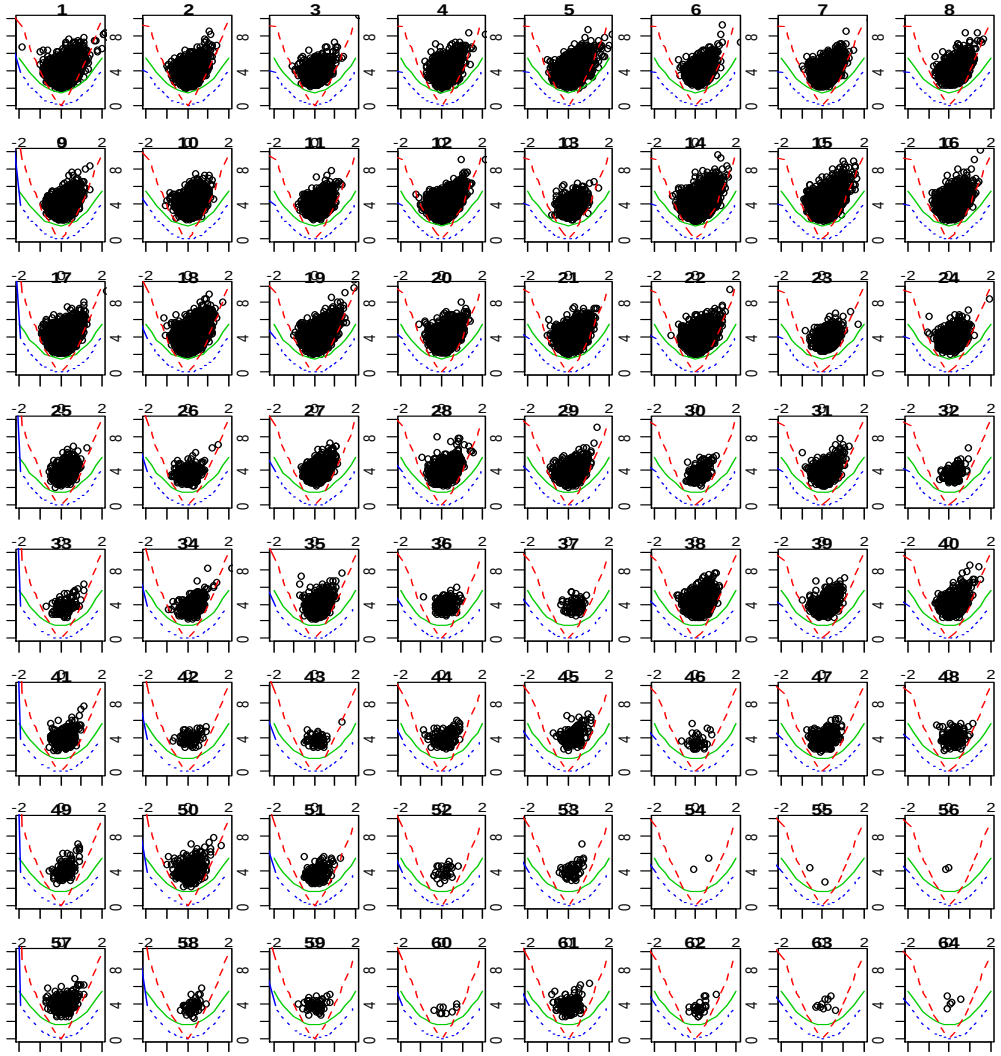
For finding any possible effect of theme relevance to aesthetic rating, all 255,530 photos in AVA dataset are classified into two groups: the free study (29,351) and the nonfree (226,179). As visualized in the S-K maps from the two groups below, any significant difference between them was not found as far as the samples in a subset are sufficient. In accordance with the report from AVA providers that relevance seems lowering consensus, relevance effect, if exist, seems not negating the found patterns (of wide kurtosis range, at least).

S-K Map of 29,351 photos from free studies in AVA dataset



(Notice: Three tags (55, 60, and 63) are absent in the free study.)

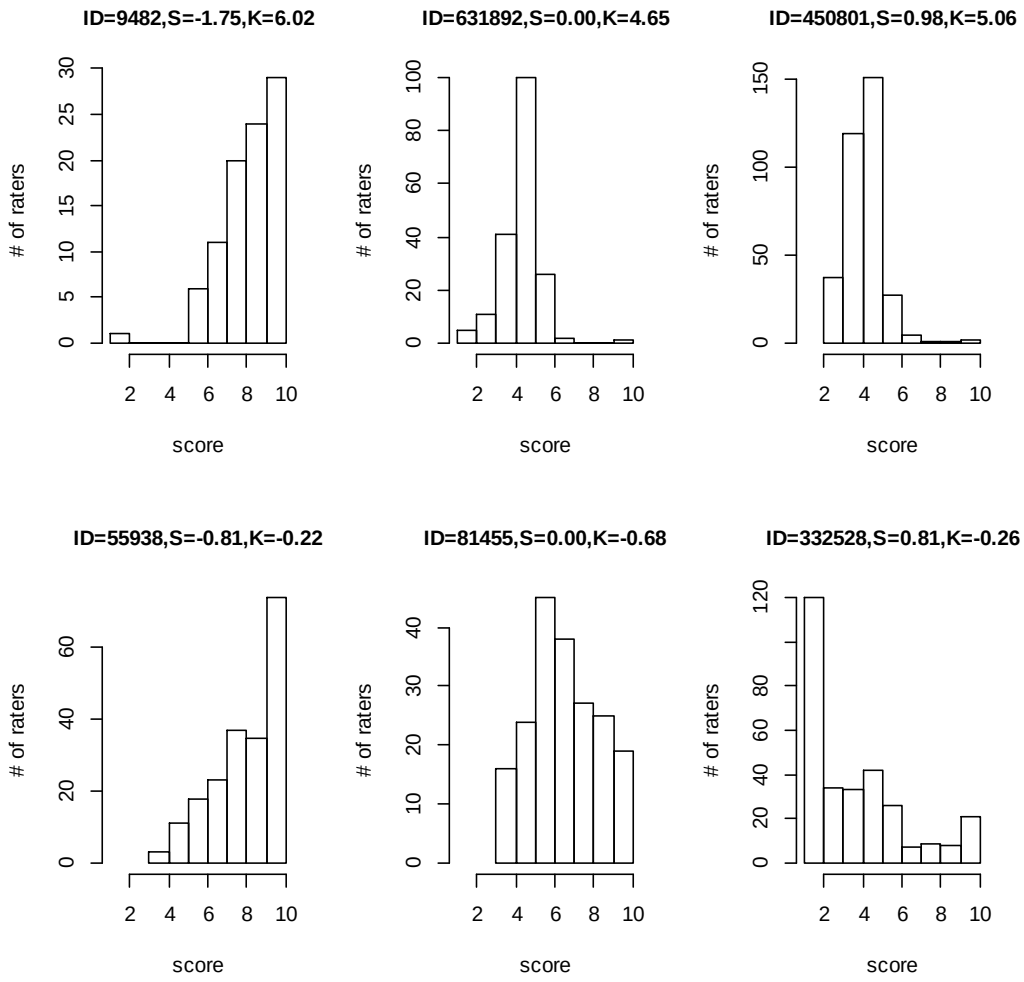
S-K Map of 226,179 photos from the non-free studies in AVA dataset



Appendix 2. Summary of Skewness and Kurtosis

Skewness is a measure of asymmetry which is usually implemented as the 3rd moment, while kurtosis has various interpretations ("peakedness" in a classical definition) on its nature and implementations including the 4th moment. In this paper, kurtosis is regarded as a measure of "lack of shoulders" and thereby captures the property of consensus among raters.

The following figure shows six representative samples from AVA dataset according to their skewness (negative for the left and positive for the right) and kurtosis (high for the upper row and low for the lower one).



Compared with the lower moments like mean or variance, skewness or kurtosis need more samples to produce valid result: hundreds, at least. Therefore, the size of target dataset does matter for consensus analysis based on S-K map.

국 문 초 록

전통적으로 철학의 영역에 속했던 “아름답다는건 무엇인가” 에 대한 탐구는 근대에 이르러 별도의 Aesthetics 분과로 정립된 인문학의 연구주제였으나, 실험미학이 태동되면서 심리학의 정량적인 측정 및 분석의 대상이 되었고 2000년대에 이르러 신경과학 및 공학에서도 관심있는 주제로 부상하고 있다. 특히 심미적 가치평가를 위한 계산모델은 자동화된 컨텐츠 추천에서 요구되는 중요한 요소로서, 주로 정지영상의 저수준피쳐에 기반한 기계학습적 접근법으로 최근 집중 연구되어왔다.

본고에서는 기존 기계학습기반 모델링에서 주로 사용된 피쳐들이 포착하기 어려운 사진내 공간/객체 표면특성이 미감에 영향을 끼칠 것이라는 가설 하에 표면특성 및 이에 기반한 2.5차원 공간배치 특성을 포착하기 위한 새로운 피쳐(LoSC)를 제안하였고, 종래 local

descriptor대비 약 30%의 연산량만으로 동등한 결과를 얻을 수 있었다. Comparative study결과는 심미적 가치평가 예측성능의 개선에 있어서 피쳐선정에 거의 영향받지 않는 일종의 유리천장 상황이 존재하고, 특히 모집단에서 절대다수를 점하는 "평범한" 샘플들이 종래의 기계학습 관점에서 난제임을 보인다.

문제의 샘플들을 정성적으로 관찰한 결과, 평가자들간의 합의수준이 크게 다른 점에 주목하여 skewness-kurtosis map을 계량화 도구로 사용하여 225,000장 규모의 사진평가 데이터셋에 대해 패턴의 보편성 여부를 검증해 보았다. 검증결과는 미적평가의 분포에 4가지 패턴들이 존재하고, 특히 합의수준의 다양성은 종래의 정상분포 가설에 잘 들어맞지 않음을 시사한다.

이에 상기 패턴들에 대한 설명력을 기준으로 다양한 심미적 가치평가 모델을 고안하여 시뮬레이션 결과를 비교검토한 결과, dynamic model 계열이 합의수준의 다양성 설명에 적합함을 확인하여 이를 근거로 심미적

가치평가 과정은 다수의 선호요인들과 혐오요인들간의 일정 시구간내 상호작용일 것이라는 가설을 세우고 이의 구현된 계산모델로서 기존 drift-diffusion model을 변형한 DDM4AP를 새롭게 제안한다.

당초 합의수준의 다양성을 설명하기 위해 제안된 상기 모델에 따르면, 합의수준과 별개로 좋거나 싫은 사진과 비교할때 평범한 사진에 대한 심미적 가치평가에 소요되는 시간이 더 길 것이라는 예측이 도출되는 바, 이를 검증하기 위해 human participant를 대상으로 수행된 실험결과에서 예측에 부합하는 latency-score간 유의미한 상관관계를 보임으로써, 상기 모델이 인간의 심미적 가치평가 과정의 주요 특성을 반영하고 있음을 시사한다.

결론적으로, 향후 기계학습기반 computational aesthetics에서 유의미한 개선을 위해선 선호요인과 혐오요인이 혼재되어 있는 상황을 고려한 학습 데이터셋 선정 및 피쳐설계, 그리고 상호대립하는 복수요인간의 동적인 상호작용 기전을 반영한 학습모델의 도입이 요구된다.

주요어 : Cognitive modeling, affect sensing and analysis, computational aesthetics, dynamic systems, consensus analysis

학 번 : 2011-30048